

# **DEEP MOLECULAR PHYLOGENY OF THE PTERYGOTA**

Von der Naturwissenschaftlichen Fakultät  
der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des Grades  
Doktorin der Naturwissenschaften  
Dr. rer. nat.  
genehmigte Dissertation  
von

Dipl.-Biol. Sabrina Simon  
geboren am 06.01.1981 in Hannover

2010

Referentin: PD Dr. Heike Hadrys  
Institut für Tierökologie und Zellbiologie  
Stiftung Tierärztliche Hochschule Hannover

Korreferent: Prof. Dr. Rob DeSalle  
Division of Invertebrate Zoology  
American Museum of Natural History

Tag der Promotion: 08.07.2010

Meinen Eltern





## Contents

<b>Zusammenfassung.....</b>	<b>1</b>
<b>Summary.....</b>	<b>3</b>
<b>1. Introduction.....</b>	<b>5</b>
1.1. Insect diversity and evolution.....	6
1.2. Insect phylogeny and hypotheses.....	7
1.3. Molecular approaches for systematics.....	11
<b>2. The aims of the thesis.....</b>	<b>13</b>
2.1. Molecular marker systems.....	13
2.1.1. Single target genes.....	13
2.1.2. Nuclear rRNA genes.....	15
2.1.3. ESTs.....	16
2.1.4. Mitochondrial genomes.....	17
2.2. Improved phylogenetic analyses.....	17
<b>3. Summary of Results and Discussion.....</b>	<b>20</b>
3.1. Single target genes.....	20
3.2. Nuclear rRNA genes.....	22
3.3. Phylogenomics.....	24
3.4. Mitogenomics.....	26
<b>4. Conclusions.....</b>	<b>28</b>
<b>5. References.....</b>	<b>30</b>
<b>6. Publications and manuscripts on which the thesis is based.....</b>	<b>38</b>
6.1. On the value of Elongation factor-1 $\alpha$ for reconstructing Pterygote insect phylogeny.....	39
6.2. Isolation of Hox cluster genes from insects in development and evolution.....	55
6.3. Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships.....	71
6.4. Comprehensive analysis of nuclear rRNA genes to infer Insect phylogeny.....	107
6.5. A phylogenomic approach to resolve the basal pterygote divergence.....	130
6.6. A phylogenomic approach to resolve the arthropod tree of life.....	156
6.7. The mitochondrial genome of two palaeopterous representatives: <i>Baetis</i> sp. (Ephemeroptera) and <i>Boyeria irene</i> (Odonata) – a mitogenomic approach to resolve the Palaeoptera problem.....	182

---

<b>7. Acknowledgements</b>	<b>205</b>
<b>8. Curriculum vitae</b>	<b>206</b>
<b>9. List of Publications</b>	<b>208</b>
<b>10. Appendix (Supplementary material associated with publications)</b>	<b>210</b>
10.1. On the value of Elongation factor-1 $\alpha$ for reconstructing Pterygote insect phylogeny	210
10.2. Isolation of Hox cluster genes from insects in development and evolution	215
10.3. Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships	217
10.4. Comprehensive analysis of nuclear rRNA genes to infer Insect phylogeny	237
10.5. A phylogenomic approach to resolve the basal pterygote divergence	242
10.6. A phylogenomic approach to resolve the arthropod tree of life	256
10.7. The mitochondrial genome of two palaeopterous representatives: <i>Baetis</i> sp. (Ephemeroptera) and <i>Boyeria irene</i> (Odonata) – a mitogenomic approach to resolve the Palaeoptera problem	280

## Zusammenfassung

### Tiefe molekulare Phylogenie der Pterygoten

Insekten repräsentieren die mit Abstand erfolgreichste Tiergruppe auf der Erde. Die Entstehung von Bauplanmodifikationen und ökologischen Anpassungen erreicht hier eine konkurrenzlose Vielfalt. Die folgenreichsten Veränderungen traten nach Erfindung der Flügel auf, also beim Übergang Apterygota (ungeflügelte Insekten) zu Pterygota (geflügelte Insekten). Offene Schlüsselfragen beinhalten den Ursprung (i) der Pterygoten (geflügelte Insekten), (ii) der Neoptera (=Neuflügler, können Flügel auf dem Abdomen ablegen) und (iii) der Holometabola (=Neuflügler mit vollständiger Metamorphose). Jedoch sind die Richtungen evolutionärer Anpassungen in wichtigen Fällen unbekannt, da die Verwandtschaftsbeziehungen zwischen den Insektenordnungen weitestgehend ungeklärt sind. Eine mögliche Ursache hierfür ist, dass die Entstehung der verschiedenen Insektenordnungen innerhalb einer kurzen Zeitspanne stattgefunden hat. Dieser Umstand zieht eine der größten Herausforderungen nach sich: das Finden von molekularen Sequenzmarkern, die diese schnelle Radiation detektieren können. Bis dato sind molekulare Datensätze für die phylogenetische Systematik der Pterygoten vor allem für abgeleitete Ordnungen generiert worden, d.h. Schlüsselfragen an der Basis der Pterygoten blieben unbeantwortet. Insbesondere Studien von Genorganisation und Komplexität beruhen auf abgeleiteten Modelorganismen wie z.B. *Drosophila*, *Apis*, *Tribolium*, *Anopheles*. Die phylogenetischen Studien, basale pterygote Insekten betreffend, basieren auf einzelnen molekularen Sequenzmarkern und erzielten widersprüchliche Ergebnisse.

Die vorliegende Arbeit hat einen kombinierten Ansatz (i) molekular-genetische Daten auf unterschiedlichen Ebenen zu erstellen und (ii) neue bioinformatische Software und Algorithmen zu testen, um die Datenanalyse zu verbessern und die Datenqualität abzuschätzen. Eine repräsentative Anzahl von allen pterygoten Insekten-Infraklassen wurde, mit speziellem Fokus auf die basalen Ordnungen, analysiert. Die Arbeit ist Teil des DFG Schwerpunktprogramms "Deep Metazoan Phylogeny" und die hier generierten Datensätze sind nicht nur neue phylogenetische Ansätze zur Analyse der Pterygoten sondern auch integrativer Bestandteil vergleichender phylogenetischer Analysen für die (i) gesamten Arthropoden und (ii) folglich der Protostomiagruppen.

Die hier neu generierten Sequenzen auf unterschiedlichen genetischen Ebenen (einzelne Target-Gene, nukleäre rRNA Gene, ESTs und komplette mitochondriale Genome) von wichtigen Insektenordnungen sind der erste Schritt für die Rekonstruktion

phylogenetischer Verwandtschaftsbeziehungen innerhalb der Pterygota und schließen die Informationslücke für entsprechende Schlüsselarten. Der methodische Ansatz wie (i) die Verwendung von neu entwickelten Algorithmen für die Identifikation von orthologen Genen (HaMStR) oder von homologen Positionen (RNAsalsa), (ii) die Implementierung von biologisch realistischen Modellen für die Sequenzevolution (Heterogenität von Abstammungslinien und einzelnen Datensätzen) und (iii) die Abschätzung von Datenqualität und phylogenetischen Signal innerhalb der einzelnen Datensätze (funktionelle Klassifizierung der Gene und der Einfluss von fehlenden Daten) führten zur besseren Auflösung an vielen Stellen des Stammbaums. Dennoch blieben manche Unstimmigkeiten bei den resultierenden Topologien bestehen, deren mögliche Ursachen hier ebenfalls analysiert wurden. Allerdings konnte für die größte Hürde in der Insektensystematik – der Ursprung der Pterygoten und somit Flügel – gezeigt werden, dass molekular-systematische Studien in der Lage sind, diese Schlüsselinnovation aufzudecken. Dieser Erfolg gibt Zuversicht für die Rekonstruktion des Insekten Stammbaums mit Hilfe von genetischer Information für Schlüsseltaxa und verbesserten phylogenetische Analysen.

Schlüsselwörter: Insekten/Pterygota, „Palaeoptera Problem“, Bauplan-Veränderungen, Molekulare Systematik, Phylogenomics

## Summary

### Deep molecular phylogeny of the Pterygota

Insects are by far the most successful animal group on Earth. Radiation of bauplan modifications and ecological adaptations have reached an unchallenged diversity in insects. The most influential insect bauplan modifications took place during and after the transition from apterygote (wingless) to pterygote (winged) insects. Key questions (and transitions) relate to: (i) the origin of Pterygota (invention of wings), (ii) the origin of Neoptera (insects with wing flexion) and (iii) the origin of Holometabola (insect with complete metamorphosis). Unfortunately, directions of evolutionary change and locations of important diversification shifts remain unresolved since deep phylogenetic relationships between insect orders are not yet resolved. One potential explanation for this situation is that these divergences took place within a limited time frame. This circumstance is accompanied by the major challenge in finding useful molecular markers to accurately track these short ancient internodes. However, the data sets available in pterygote molecular systematics have a strong bias towards the derived orders of insects. Especially studies of gene organization and complexity mostly focus on particular model systems such as *Drosophila*, *Apis*, *Tribolium*, *Anopheles* etc.; and most phylogenetic studies on basal pterygote insects are based on individual molecular sequence markers and reveal conflicting results.

In this thesis, (i) molecular genetic data at different levels of genetic and genomic organization were compiled and (ii) new bioinformatic tools and algorithms were applied to improve data analyses and data quality assessment. A representative sampling of all pterygote infraclasses were analyzed with a special focus on the most basal pterygote orders. The thesis is part of the DFG special priority program “Deep Metazoan Phylogeny”. All data sets were generated not only for new phylogenetic approaches to infer pterygote relationships, but also as an integrative part of comparative phylogenetic studies in the (i) Arthropoda and (ii) consequently the Protostomia lineages.

The new sequences at different genetic levels (single target genes, nuclear rRNA genes, ESTs and complete mitochondrial genomes) from several crucial insect orders provided a first step towards resolving the phylogenetic relationships of Pterygota and closing the gap of information for key taxa. The performed methodological approaches such as (i) the application of new developed algorithms for the assignment of orthologous genes (HaMStR) or homologous positions (RNAsalsa), (ii) the implementation of

biologically realistic models of sequence evolution (heterogeneity among lineages and data partitions) and (iii) the assessment of data quality and phylogenetic signal within the data partitions (functional classification of genes and impact of missing data) resulted in better resolution in several places, but some incongruence among the inferred topologies still remained. However, for the major obstacle in insect systematics – the question of the origin of Pterygota and consequently of the invention of insect wings – it is shown that molecular systematic approaches are able to unravel this key evolutionary innovation and give further confidence for the reconstruction of the Insect Tree of Life by closing the gap of genetic information and improved phylogenetic analyses.

**Key words:** Insects/Pterygota, “Palaeoptera Problem”, bauplan transitions, molecular systematics, phylogenomics

## 1. Introduction

This thesis was a combined approach (i) to compile molecular genetic data at different levels of genetic and genomic organization and (ii) to develop and apply new bioinformatic tools and algorithms to improve data analyses and – most important – data quality assessment. A representative sampling of all pterygote infraclases was analyzed with a special focus on the most basal pterygote orders. The thesis was part of the DFG special priority program “Deep Metazoan Phylogeny” (SPP1174) (DMPP hereafter) (<http://www.deep-phylogeny.org>). All data sets of this thesis were generated not only to solve pterygote phylogeny but also as an integrative part of comparative phylogenetic studies in the (i) Arthropoda and (ii) consequently the Protostomia lineages.

Within the DMPP phylogenetic relationships of Metazoa, including the whole spectrum from basal diploblasts to derived Protostomia and Deuterostomia, are being investigated. The DMPP started in August 2005 and 32 projects are funded in total over a period of 6 years. The projects comprise a variety of new morphological, molecular and bioinformatic approaches across the metazoan kingdom. Besides the primary aim of the DMPP to compile a mass of molecular and morphological data in order to reconstruct a robust backbone tree of metazoan life, the long-term goal focuses on detecting the direction of major evolutionary events that gave rise to distinctive taxonomic signatures of metazoan life. The thesis is “located at the tip of the Arthropoda” and covers the most diverse and species rich group, the Pterygota (winged insects) with special emphasis on the apterygote-apterygote transition. It is closely interwoven with three other DMP arthropod projects, covering the basal hexapod lineages “Apterygota”, the Crustacea and the Chelicerata and further provides data for the overall bioinformatic projects on EST analyses and mitochondrial genome analyses, e.g. Simon et al. 2009 (see 6.5), von Reumont et al. 2009 (see 6.3) and Meusemann et al. 2010 (see 6.6).

### 1.1 Insect diversity and evolution

With respect to diversity in form, life history and adaptation, species and taxa richness, insects have an unassailable lead. They compose the bulk of arthropod species and are amongst the most successful animal groups of our planet, representing approximately 60% of the diversity of life [1] (Figure 1). There are only around 1 million taxonomically described insect species but the expected number of species ranges between 2.5 million and an incredibly species richness of 30 million [1, 2], but see discussion in [3].

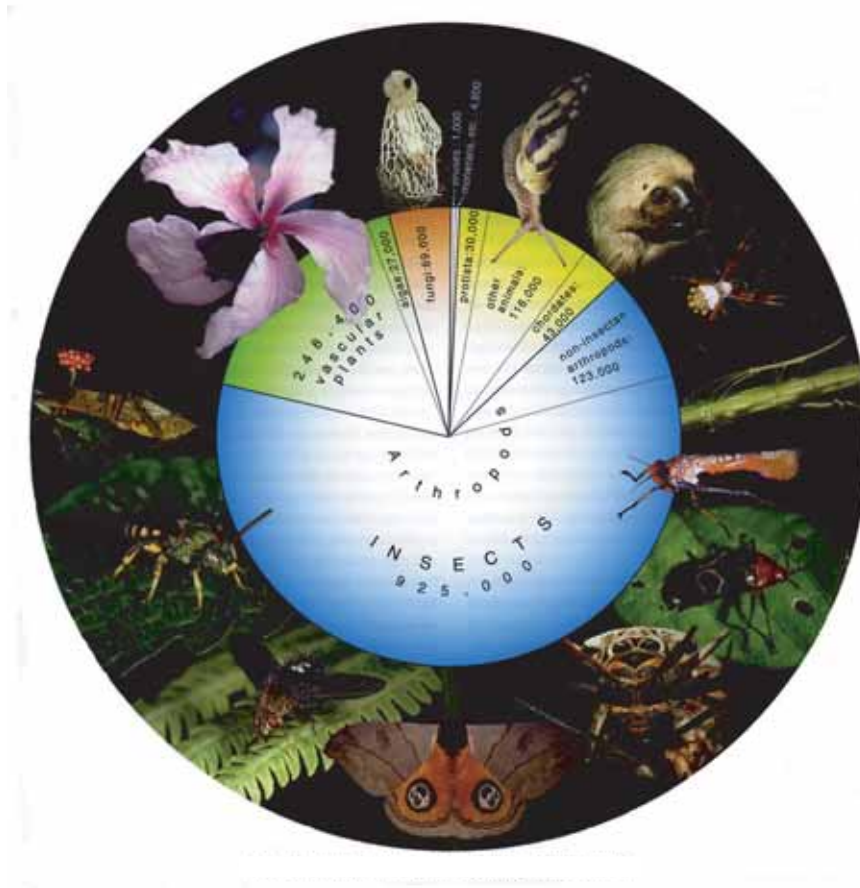


Figure 1: The diversity of life shown as proportion of named species, after [1].

Insects have also one of the widest distributions on Earth and occupy any imaginable niche, from the desert (up to 70°C: the Sahara Desert ant *Cataglyphis bicolor* [4]) to the artic (down to -60°C: larvae of the fly *Heleomyza borealis*, [5]) and even the world's deepest lake Baikal in the Siberian region of Russia is inhabited by insects (down to 1,360 m: larvae of the midge *Sergentia koschowi* [6]).

The enormous diversity of insects is linked to the evolutionary processes and innovations which occurred during their evolution. The theory of evolution unites all fields of biology and every biological phenomenon can be explained in an evolutionary context – from behavior to speciation processes to gene mutations. Charles Darwin (1809-1882)



proposed a theory of evolution by means of natural selection at a meeting of the Linnean Society of London (1858). His monumental treatise “On the Origin of Species” published one year later still is the basis of modern evolutionary theory to explain the diversity of life. He proposed that all species of life have evolved over time from one or a few common ancestors through the process of natural selection and sketched a genealogical branching of a single evolutionary tree – the Tree of Life.

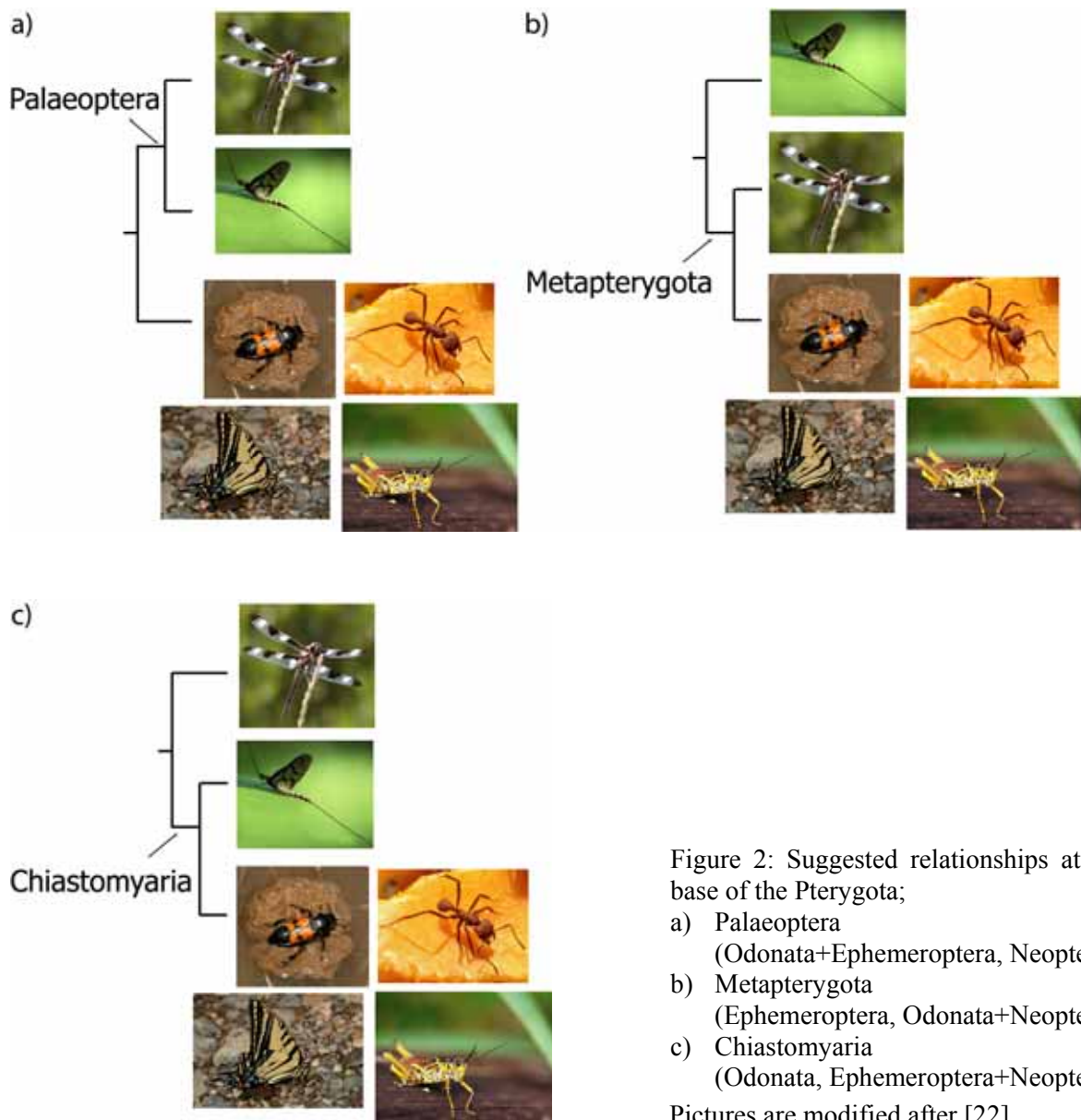
Unraveling the Insect Tree of Life could provide essential insights into the evolution of global biodiversity, evolutionary processes and key innovations. In insect evolution at least four events are commonly cited as being of major macro-evolutionary importance: the origin of (i) the insect bauplan, (ii) the Pterygota (winged insects), (iii) the Neoptera (insects with wing flexion) and (iv) the Holometabola (insects with complete metamorphosis), [7-13]. These unique features are implicated as major contributors to the vast insect diversification, since they allowed the accessibility of ecological niches (insect bauplan), dispersal ability (wings), exploration of concealed places (folding of wings) and specialized juvenile and adult forms (complete metamorphosis), see [14]. However, none of them could be identified unambiguously as important shifts in diversification due to lacking phylogenetic information [13, 15].

## 1.2 Insect phylogeny and hypotheses

Systematically, insects are the most studied animal class and are also often used as model systems in evolutionary developmental biology (e.g. *Drosophila*, *Anopheles*, *Tribolium*). This fact is not surprising regarding that Willi Hennig (1913-1976), the founder of phylogenetic systematics, was an entomologist specialized in dipterans. In insect systematics whether based on morphological or molecular data, numerous characters have been assembled and their states have been interpreted to unravel the evolutionary history of insects.

Undoubtedly, the invention of wings is one of the most important key innovations in insect evolution since wings allowed the conquest of a large number of new ecological niches and improved the capability of dispersal and evasions of predators [1, 8]. But at present, neither the evolutionary origin of the insect wings (the origin of Pterygota from apterygote insects) nor their radiation into a broad spectrum of sophisticated flight organs are well understood. For example, combined analyses of extensive morphological and molecular data by Giribet et al. [16] places one of the basal pterygote orders, the Ephemeroptera (mayflies) in a sister group relationship to the “apterygote” order

*Zygentoma* (silverfish). The latter is characteristic for our current understanding of deep branching patterns in insects, particularly for the most central ones, the relationships at the base of the Pterygota (Figure 2). Here the three branches Ephemeroptera (mayflies), Odonata (dragonflies and damselflies) and Neoptera (all other extant insects with wing flexion) compete in three different scenarios: (i) the traditional scenario – the “Palaeoptera hypothesis” – suggests a monophyletic Palaeoptera clade, including Ephemeroptera and Odonata [17]; (ii) the Metapterygota hypothesis – a clade formed by Odonata and Neoptera – places Ephemeroptera at the base of the pterygotes [18-20] and (iii) the Chiasmomyaria hypothesis – a clade formed by Ephemeroptera and Neoptera – places Odonata at the base of the pterygotes [21].



The origin of Neoptera also deserves particular attention. Mayhew [13] considers the Neoptera (Polyneoptera, Paraneoptera and Holometabola) to represent the most significant insect radiation. The phylogenetic relationships of the Neoptera are highly problematic and the morphological synapomorphies defining the polyneopterous orders at the base of the Neoptera remain unresolved. The polyneopteran monophyly is not generally accepted and the relationships within the Polyneoptera are only partially clear. Especially the placement of the Plecoptera, Dermaptera and the highly specialized Zoraptera have been key issues of debate for many years since each of them is without clear affinity to any other polyneopterous order and each one is a recent remnant of an otherwise ancient lineage [1]. Kristensen [20] pointed out that the relationships among basal neopteran orders are an nearly unresolved “comb”.

Another potential important shift in hexapod diversification occurred at the origin of the Eumetabola (Paraneoptera+Holometabola; insects with complete metamorphosis plus bugs and their relatives), an often neglected event in insect evolution [13]. However, there is still no consensus about the Eumetabola hypothesis. This group achieves high support through morphological characters, like the development of the male genital structures [21], but so far, none of the cladograms based on molecular data supports the Eumetabola hypothesis [18, 19, 23].

The same holds true for the origin of Holometabola (insects with complete metamorphosis). Neither morphological nor paleontological or molecular data could unequivocally resolve the origin and radiation of holometabolous insects [24-27]. It will be shown in the future if the results of Savard et al. [28] and Wiegmann et al. [29] – Hymenoptera as the earliest branching holometabolan lineage – can be confirmed.

Despite the large number of hypotheses we are not even close to resolving the origin of the most influential shifts in insect evolution. The main problem here simply relates to a widely unresolved insect phylogeny in general and the gap of data in the basal part of the pterygote tree [30]. Several reasons have been proposed why it is so difficult to recover the evolutionary tree of insects. Each of them accounts for the problems in insect systematics: (i) fast-evolving and enormous diversity of insects (“ancient rapid radiation”), (ii) preserved ancient characters in some taxa and (iii) lack of fossil records.

According to molecular clock analyses insects arose in the Late Silurian (approx. 420 million years ago (MYA)) coinciding with the earliest plant megafossil [31]. Engel and Grimaldi [32] confirmed this origin of insects based on fossil records and provided further evidence for an origin of pterygotes in the mid-Devonian (approx. 387 MYA); see



### 1.3 Molecular approaches for systematics

Molecular phylogenetic studies require appropriate molecular marker systems with phylogenetically informative sites tracking the window, within which the divergence and origin of lineages took place. The evolution of insects occurred in the ancient past (~420 MYA) within a short period of time followed by a rapid radiation into a vast number of species (see also Figure 3). Whitfield and Kjer [33] pointed out that this “ancient rapid radiation” causes the main problem when resolving insect relationships using molecular data. Due to short ancient internodes connecting the taxonomic groups, inadequate molecular data sets and conflicting results within or among datasets have been observed in many insect phylogenetic studies [18, 23, 34, 35].

Consequently, a major challenge is to find useful molecular markers to accurately track these short ancient internodes, which have kept pace with speciation, but slow enough to have transferred the phylogenetic signal to the present [36]. Unfortunately, the rationale behind the selection of certain molecular markers is not always clear, leading to discrepancies and incongruence between individual gene trees and unresolved phylogenetic trees [18, 34]. Furthermore the preserved ancient characters in some taxa and the rate heterogeneity among orders lead to confusion among phylogeneticists. For example, Kjer et al. [34] observed excessive substitution rate acceleration in rRNA data for Diptera and Diplura, while Odonata and Mantodea seem to almost “stand still”.

In this context, phylogenetic trees inferred from 'supermatrix' analyses or the 'supertree' approach have become a popular method to resolve long outstanding questions in deep metazoan relationships, e.g. [37-42]. There is still a controversy about phylogenetic reconstructions derived from 'supertree' versus 'supermatrix' approaches and both methods have demonstrated their strengths and weaknesses [43-46]. However, some promising new approaches address the existing problems. For the 'supertree' method, for example, the recent proposal of a maximum likelihood approach introduces an important idea for future phylogenetic inferences from genomic data [47, 48]. On the other hand, the capabilities of supermatrix analyses have been increased by the recent development of new tree reconstruction methods, including model-based techniques for analyzing heterogeneous data and hierarchical methods for constructing extremely large trees; reviewed in [49]. By applying more realistic evolutionary models for tree reconstruction, systematic errors could be overcome [50, 51]. A study of Lartillot et al. [52] has shown that, e.g. the use of site-heterogeneous models is more robust against long branch attraction (LBA) artifacts.

Recent discussions center also around the impact of gaps or equally missing data on large alignments for reconstructing the “accurate” topology. Hartmann and Vision [53] have shown that EST-like gappiness can pose a major problem for phylogenetic analyses. Other studies have shown that missing data, in terms of incomplete taxa sampling, are not necessarily problematic as long as the overall number of phylogenetic informative characters is high [27, 54-56]. In turn, Dunn et al. [38] improved the overall bootstrap support by excluding unstable taxa which were calculated by leaf stability indices where some of these unstable taxa were identified due to the poor gene sampling. Also the identification of orthologous genes remains a challenge. In this context several researchers suggest using ribosomal proteins as molecular markers due to their abundance in EST data sets and the non occurrence of paralogs within Metazoa [27, 37, 39, 40]. However, Dunn et al. [38] pointed out that ribosomal proteins should be carefully evaluated for paralogy, the same as for any other type of gene. Consequently, the above mentioned studies are questionable due to paralogy problems which can lead to systematic errors.

It is also observed that in some problematic cases, resolving deeper phylogenetic relationships by simply increasing the number of nuclear sequence markers seems to fail. Thus, an increasing number of systematists believe that in many of those cases the inclusion of not only genome data but also data on gene organization and complexity as well as new methodical approaches to assess data quality becomes essential, e.g. [42, 57, 58]. Eventually comparative research on regulatory genes may become helpful for deep phylogenetic studies and might bridge some gaps between description and causal explanations.

Although molecular data are not the panacea for an “accurate” resolution of the insect tree, molecular data, with appropriate choice of taxa and data, help to resolve certain phylogenetic questions that morphology has been unable to answer. Especially for deeper phylogenetic relationships (family or ordinal level) where it might be difficult or even impossible to establish homologous character states in compared taxa due to differences in appearance and function, molecular markers are an indispensable source of information for systematics.

## **2. The aims of the thesis**

The aim of this thesis is to provide a variety of approaches for the reconstruction of the insect evolutionary tree by analyses of molecular data at different genetic levels. At the conceptual level data was obtained by four means: (i) characterizing new target genes of importance at different levels of gene and bauplan organization, (ii) characterizing and analyzing rRNA data by incorporating knowledge of secondary structure predictions, (iii) compiling extensive EST data and evaluating phylogenetic informative partitions within the dataset and (iv) retrieving new phylogenetic information from mitochondrial genome analyses.

At the methodological level new algorithms and software tools were tested by using the above data sets, e.g. identification of gene orthology (HaMStR [59]), alignment algorithms (RNAsalsa [60]) and software for data quality assessment prior to tree reconstructions (ALISCORE [61], MARE (Misof et al., unpubl.)).

### **2.1 Molecular marker systems**

#### **2.1.1 Single target genes**

The most commonly used molecular markers systems to infer relationships between pterygote orders have been single mitochondrial genes and nuclear rRNA gene fragments. However, both have shown severe limitations for phylogenetic analyses in Pterygota [25, 35, 62, 63]. The rapid evolution of mitochondrial coding genes makes them useful for studies at intraordinal levels, but they are often ill-suited for resolving deeper nodes [64]. Although ribosomal RNA genes obviate the above problem, assessment of positional homology becomes a problem due to the occurrence of indels (see also 2.1.2). The evaluation of new markers, especially nuclear protein coding genes, has become a necessity for resolving pterygote phylogeny. Nuclear protein coding genes evolve at a slower rate than mitochondrial protein coding genes and show little length variability and consequently obviate the above described problems of other molecular markers. Apart from Histone 3 (H3 hereafter) – a nuclear protein coding gene frequently used in insect systematics – some additional nuclear protein coding genes have been used more often, including the Elongation factor-1 $\alpha$  (EF-1 $\alpha$  hereafter). EF-1 $\alpha$  which is a relatively conserved nuclear gene facilitating GTP dependent binding of tRNAs to the acceptor site of ribosomes [65], has been proved useful in some terrestrial arthropods [30, 66, 67]. The variation in silent nucleotide sites makes this marker useful for phylogenetic analyses

among species groups and genera [68-71], while amino acid replacements provide phylogenetic information for deeper divergences [36, 72].

In order to evaluate the phylogenetic value of both nuclear protein coding genes, Elongation factor 1- $\alpha$  and Histone 3, these genes have been isolated and characterized for a representative sampling of pterygote orders. Statistical and phylogenetic analyses for both genes were conducted and “structural” features in terms of intron positions within EF-1 $\alpha$  were identified which might provide additional phylogenetic information (Simon et al. 2010, see 6.1).

Another set of single target genes are regulatory genes which have often been discussed in a phylogenetic context and when studying the complicated evolutionary history of insects. The explanatory power of these genes in developmental research is without question and their genomic organization has also supported important clades in metazoan phylogeny [73, 74]. At the sequence level these genes may also harbor some phylogenetic information. There is – for example – strong evidence that changes in the sequence and expression (function) of the homeotic genes *ultrabithorax* (*Ubx*) and *sex combs reduced* (*Scr*) in basal pterygote orders relate to various stages of abdominal appendages [75, 76]. In the derived holometabolous insects, *Ubx* relates directly to the evolution of morphological diversity in hindwings. Comparative sequence and functional studies of these genes may shed some light on the early evolution of insect flight, the most important key innovation in hexapod diversification. However, no comparative data exist for basal insects and no Hox genes have been isolated from the majority of insect orders yet.

Regulatory genes (Antennapedia (*Ant*) class genes) have been isolated and characterized for 6 basal insect orders (“apterygote” and pterygote species; *Campodea fragilis* (Diplura), *Lepismachilis y-signata* (Archaeognatha), *Sympetrum sanguineum* and *Ischnura elegans* (both Odonata), *Baetis* sp. (Ephemeroptera), *Nemoura cinerea* (Plecoptera) and *Forficula auricularia* (Dermaptera)) (Hadrys et al. 2010, see 6.2). The phylogenetic analyses were performed on a concatenated matrix from nine homeodomain sequences, and the results indicate that regulatory sequences may harbor some phylogenetic information.



### 2.1.2 Nuclear rRNA genes

The structural complexity of rRNA molecules – highly variable expansion segments (length and sequence heterogeneity) between conserved and slowly evolving core regions – challenges the assignment of homologous positions within alignments. The variable regions are difficult to align across diverse taxa and are consequently usually omitted in tree reconstruction [77]. Moreover, the site heterogeneity of evolutionary rates, the non-stationarity of base composition among taxa and rate variation in time observed in rRNA molecules are the main problems for reconstructing ancient splits due to the erosion of the phylogenetic signal (see also 2.2).

The topology of a phylogenetic tree may be critically dependent on the accuracy of the sequence alignment employed [78, 79]. In order to increase the dependability of rRNA sequence alignments the secondary structure of the ribosomal RNA were taken into account as a guide for the assignment of positional homology by using RNAsalsa [60]. The secondary structure of rRNA molecules is highly conserved throughout evolution despite primary sequence divergence due to the physiological function of ribosomal RNA in protein biosynthesis mostly maintained by the molecule's secondary and tertiary structures [80, 81]. The incorporation of existing knowledge on secondary structure has been proven to be promising source for increasing alignment accuracy [23, 82, 83]. Moreover, the inferred consensus secondary structure and the individual structure predictions can be used (i) to incorporate the site covariation in the phylogenetic analyses and (ii) to evaluate the phylogenetic signal in the structure of rRNA molecules.

Mainly 18S rRNA genes and fragments of 28S rRNA genes were used to infer insect phylogeny, with contradicting results regarding the basal pterygote divergence. Hovmöller et al. [84] supported the Palaeoptera hypothesis using the complete 18S rRNA sequence and a fragment of the 28S rRNA. The Metapterygota hypothesis is supported by Wheeler et al. and Ogden and Whiting [18, 35] using fragments of 18S and 28S rRNA. The Chiasmomyaria hypothesis is recovered using the complete 18S rRNA [23]. Mallat and Giribet [85] also supported the Chiasmomyaria hypothesis using nearly complete 18S and 28S rRNA genes. However, their data set is characterized through the underrepresented insect/pterygote species (8 apterygote insects and 15 pterygote insects), due to the lack of sequence information mainly for the complete 28S rRNA genes. For pterygote insects the complete 28S rRNA gene has been only characterized for 21 pterygote species (12 orders). New pterygote specific primers were developed and the complete 28S rRNA gene was amplified across a broad range of pterygote species which were used for (i) phylogenetic

analyses of major arthropod lineages by incorporating background knowledge of ribosomal RNA data (von Reumont et al. 2009, see 6.2) and (ii) phylogenetic analyses at the sequence and secondary structure level of insects (Simon & Hadrys 2010a, see 6.3).

### 2.1.3 ESTs

ESTs (Expressed Sequence Tag) are short DNA sequences corresponding to a fragment of a complementary DNA (cDNA) molecule and which is expressed in a cell at a particular given time. They provide a comprehensive random sample of protein coding genes and an economic way to produce large amounts of sequences for phylogenetic analysis of “non-model” species for which genome sequence projects are not yet available. While there is still a limited explanatory power of the standard marker genes used in insect phylogenies, the use of ESTs is a straightforward quantitative approach to increase the number of characters and improve the explanatory power of sequence analyses.

In addition, multi-gene analyses (phylogenomics), derived from genome- and EST-projects, are the state-of-the-art approaches to resolve deep metazoan relationships, see [28, 38, 86-90]. However, comparative EST analyses with a phylogenetic background for insects are still rare and ESTs from basal Pterygota are non existent; see [87, 91, 92]. Focusing on the gap of information at the base of the Pterygota, EST projects were conducted on Odonata (*Ischnura elegans*) and Ephemeroptera (*Baetis* sp.) in order to (i) address the “Palaeoptera problem” by using a phylogenomic approach (Simon et al. 2009) and (ii) to reconstruct a backbone tree of arthropods (Meusemann et al. 2010).

Besides the primary aim of both studies to reconstruct a robust phylogeny, the two studies aimed to assess the data quality and the phylogenetic signal in different approaches. In the study of Meusemann et al. (2010) an optimal data set was selected by the application of newly developed bioinformatic tools (MARE; Misof et al., unpubl.) to increase the number of taxa with potentially informative genes, by excluding poorly represented taxa and uninformative genes. This approach should improve the signal-to-noise ratio in the data and reduce the effort spent in tree reconstructions (see 6.5).

To evaluate the phylogenetic content in the EST data set another approach was performed for the phylogenomic approach resolving basal pterygote relationships (Simon et al. 2009, see 6.4). Considering that different evolutionary signals are a result of the different evolutionary processes that act upon the genes and that the functional role of these genes in the cell is important for the phylogenetic signal they carry [93], the biological function of the genes in the data set was assessed. Statistical analyses on the subdivided

data set – according to their function – were performed to analyze the phylogenetic value of the genes for resolving the “Palaeoptera Problem”.

#### 2.1.4 Mitochondrial genomes

Up to March 2010, approximately 170 complete mitochondrial genomes for insects are available from the GenBank nonredundant database, within a strong bias to derived orders. 149 genomes for Neopteran species were available (29 polyneopterans, 44 paraneopterans and 76 holometabolans) with an over sampling of species of the some order. But still some representatives of insect orders are underrepresented or even missing (e.g. Protura, Odonata, Ephemeroptera, Dermaptera, Zoraptera, Plecoptera, Embioptera, etc.).

Zhang et al. [94] published the first complete mitochondrial genome of a mayfly and conducted a mitogenomic approach to resolve basal pterygote divergence. Until now there are three mt genomes of odonates (*Orthetrum triangulare* (nearly complete) [95], *Davidius lunatus* [96] and *Pseudolestes mirabilis* (FJ606784; unpubl.)) and three mt genomes of mayflies (*Parafronurus youi* [94], *Ephemera orientalis* [96], *Siphonurus immanis* (NC\_013822; unpubl.)) available. So far no combined primary sequence analyses for phylogenetic inferences have been performed on the complete data set. Regarding the mitogenomic approach of Zhang et al. [94], the data set supported the Metapterygota hypothesis, but the used data set could also be influenced by sparse taxon sampling (only one odonate and one mayfly is represented). To improve the taxon sampling of the basal pterygote orders two crucial palaeopterous mitochondrial genomes were added to the existing data set (the dragonfly *Boyeria irene* and the mayfly *Baetis* sp.). Furthermore a mitogenomic approach was applied to resolve the “Palaeoptera Problem”.

All mt genome datasets developed with by DFG special priority program are also used for combined analyses to resolve the Metazoan Tree of Life.

#### 2.2 Improved phylogenetic analyses

Although molecular data offer a promising approach to resolve certain phylogenetic questions that morphology has been unable to answer due to unrecognized homologous character states, they still sometimes fail due to strength and pitfalls of the data and the applied analyses. Major problems concern (i) establishment of positional homology in the alignment (due to sequence length heterogeneity), (ii) LBA (“long-branch attraction”) effects (due to rapidly changing lineages), (iii) loss of phylogenetic signal (due to

convergent or parallel substitutions), (iv) identification of orthologous sequences (due to gene duplication) and (v) choice of substitution models for the phylogenetic analyses (due to lineage specific variation of the model of evolution and compositional heterogeneity of base frequency); see [50, 97-101].

With the aim to improve the phylogenetic analyses of rRNA data, new bioinformatic tools (RNAsalsa [102] and ALIScore [61]) were applied for quality control of data prior to tree reconstruction. RNAsalsa combines structural and primary rRNA sequence alignment criteria and predicts ribosomal RNA folding [60]. The inferred site covariation patterns can subsequently be used to guide the application of mixed substitution models in the phylogenetic analyses. In addition, (i) biologically realistic mixed DNA/RNA substitution models in a Bayesian approach and (ii) lineage specific variation of the model of evolution (time-heterogeneous) were applied using the software PHASE 2.0 [103].

Assignment of gene orthology is essential when using ESTs as a source for phylogenetic analyses due to the complex nature of genome evolution, involving gene loss, duplications, expansion of gene families and functional diversification. To assign gene orthology the new search algorithm HaMStR [59] (Hidden Markov Model Based Search for Orthologs using Reciprocity, developed within the DMPP) was applied which combines a profile Hidden Markov Model search and a subsequent BLAST search. The data quality was further assessed since it currently appears that adding extensive new data to multi-gene approaches is not always the method of choice in difficult groups, apparently due to the effect of certain systematic biases [50, 104]. LBA effects, for example, seem to increase when new data are added to difficult groups. The Pterygota represents the largest group within the metazoan tree of life. They are an excellent group to assess data quality in general (checking information content of single genes, checking for split conflict and potential LBA effects and random similarities within alignments) and for characterization of individual genes and genomic regions for their congruence with overall evolutionary history.

Mitochondrial genomes evolve at higher rates than the nuclear genome [93], and can cause saturation of the phylogenetic signal which can be problematic in deep split phylogenies. The phylogenetic signal becomes noisy due to multiple substitution processes (saturation) and results in random similarity of alignment regions. Such noisy sections potentially bias tree reconstructions in several ways and exclusion of these saturated regions can help to reduce noise [105, 106]. In order to identify saturated regions within

the alignment ALISCORE was applied. ALISCORE generates profiles of randomness caused by alignment ambiguity or saturation using a sliding window approach [105]. In addition, to address the problems in the analyses of mt-genome data, particularly the effect of long-branch attraction (LBA), appropriate models that accommodate rate heterogeneity across data partitions were applied [107, 108].

### 3. Summary of Results and Discussion

(von Reumont et al. 2009; Simon et al. 2009; Simon et al. 2010; Meusemann et al. 2010, Hadrys et al. 2010, Simon & Hadrys 2010a, Simon & Hadrys 2010b)

In the studies upon which this thesis is based, various molecular marker systems at different genetic levels were isolated and explored to infer pterygote phylogeny and in a grander scheme arthropod phylogeny (see 6.3 and 6.6). A representative sampling of all pterygote infraclases were analyzed with a special focus on the most basal orders due to the still existing gap of sequence information.

Methodological innovations developed in the DMPP by means of new algorithms and data quality assessment tools were applied, tested and improved accordingly to the data partitions, e.g. identification of gene orthology (HaMStR [59]), alignment algorithms (RNAsalsa [102]) and software for data quality assessment prior to tree reconstructions (e.g. ALISCORE [61], MARE (Misof et al., unpubl)). The present studies improved the phylogenetic reconstruction methods and evaluated methods for data quality assessment which resulted into a better resolution in several places of the inferred topologies, e.g. Chiasmomyria hypotheses supported by rRNA data and ESTs.

#### 3.1 Single target genes

(Simon et al. 2010, Hadrys et al. 2010 and references therein)

Comparative phylogenetic and statistical analyses of the Elongation factor 1- $\alpha$  and Histone 3 highlighted the superiority of EF-1 $\alpha$  over Histone H3. Possible saturation of substitutions was evaluated for both genes by plotting the observed number of transitions (ti) and transversions (tv) at each codon position against uncorrected (“ $p$ ”) sequence divergence. In addition the rate matrix, transition/transversion ratio, proportion of invariable sites and  $\alpha$  shape parameter were estimated. Maximum likelihood (ML) and Bayesian analyses (BI) for EF-1 $\alpha$  and Bayesian analyses for Histone 3 were performed.

The results indicate that Histone 3 should be replaced and that EF-1 $\alpha$  should be used as an additional informative molecular marker in pterygote systematics, as previously suggested by Kjer et al. [109]. The statistical analyses revealed that Histone H3 shows extreme conservation in the second position and substitutions occur mainly at silent sites, whereas in EF-1 $\alpha$  substitutions occur more frequently. The phylogenetic analyses showed that a small exon fragment of EF-1 $\alpha$  provides a topology separating the infraclases which

corresponds to evidence from other molecular markers and morphology, while the Histone 3 topology has no resolution at these deep nodes.

In addition analyses of intron positions of the amplified EF-1 $\alpha$  fragment revealed same potential “structural” characters. Other studies demonstrated that these “structural” characters may mirror relationships within insect orders and that recent intron insertions can characterize monophyletic groups [110, 111]. The identification of an intron which is present in almost all pterygote orders but is absent in the derived holometabolous orders, indicates an intron loss in the evolution of the Holometabola.

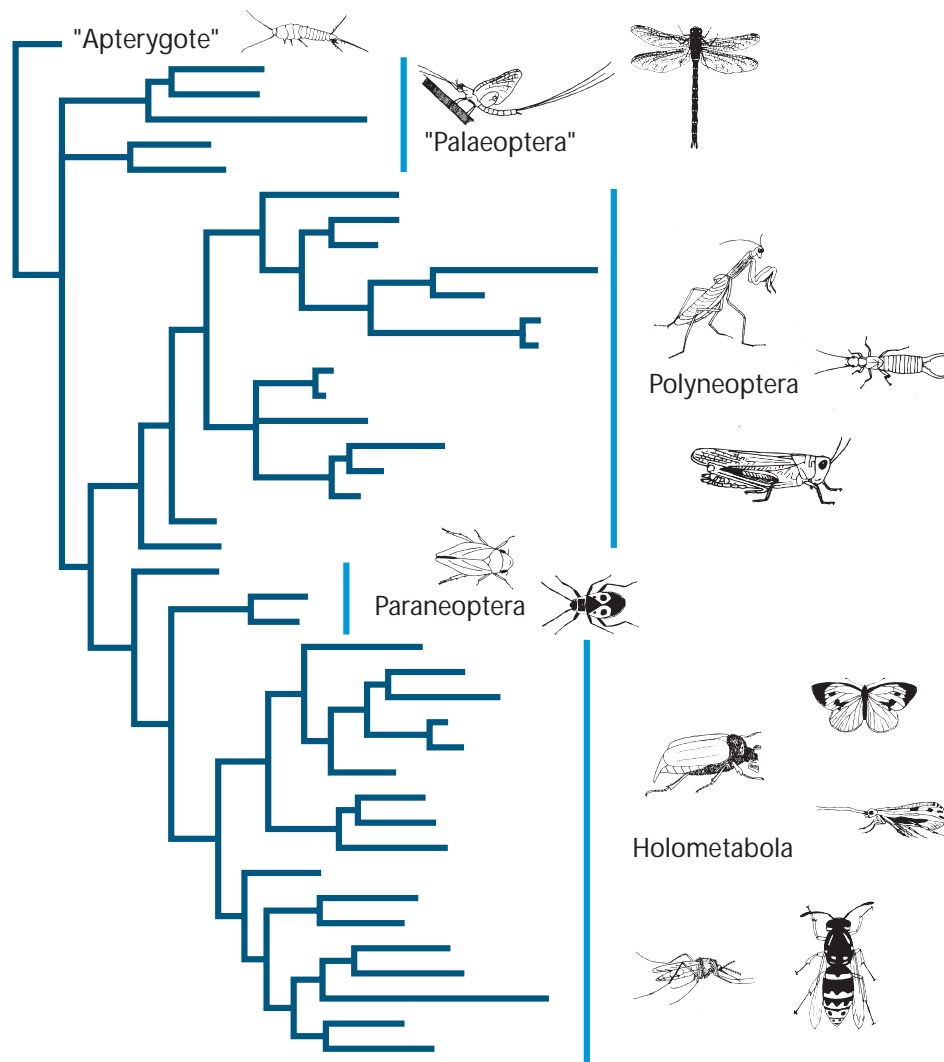


Figure 4: Topology based on EF-1  $\alpha$ , modified from Simon et al. 2010, see also 6.1. Pictures are modified after [112].

Regulatory genes (Antennapedia (*Ant*) class genes) code for homeodomain-containing transcription factors and are of outstanding importance for metazoan radiation [e.g. 113]. These genes provided deep insights into both, the phylogenetic patterns and the genetic mechanisms of animal bauplan development, consequently representing a direct bridge to

morphology [e.g. 114, 115-117]. Marden et al. [118] highlight the crucial importance of sequence information for these genes from basal Pterygota in order to reveal intermediate stages of evolution of appendages and shed some light on the early evolution of flying insects. Therefore 35 new partial homeobox sequences of *Antp*-class genes (*lab*, *pb*, *Dfd*, *Scr*, *ftz*, *Antp*, *Ubx*, *abd-A* and *Abd-B*) were isolated for basal “apterygote” and pterygote orders. Preliminary phylogenetic analyses of a concatenated matrix from nine homeodomain sequences support a basal position of Ephemeroptera but also place major Holometabola groups between basal infraclases. This is most likely the result of the large lack of sequence information from the lower pterygote orders and the amount of missing data for several homeodomains of different taxa. However, the preliminary tree topology indicates that regulatory sequences may harbor some phylogenetic information, e.g. the separation of apterygote and pterygote clades and the support for intraordinal relationships.

### 3.2 Nuclear rRNA genes

(von Reumont et al. 2009, Simon & Hadrys 2010a and references therein)

For the Pterygota, insect specific primers were designed for sequencing of the complete 28S rRNA gene including the rarely sequenced 3' end of the gene, the expansion segment D12. This region has been only explored for five insect species yet (*Aedes albopictus* [119], *D. melanogaster* [120] others therein), *Acyrtosiphon pisum* [121], *Tenebrio* sp. [122] and *Apis mellifera* [123]. The complete 28S rRNA gene for 71 pterygote species from 28 orders has been sequenced for this thesis. From this data set 26 sequences were used for the study inferring arthropod phylogeny (see 6.3).

Results of the cooperative work (see 6.3) show that the implementation of biologically realistic model parameters, such as site interaction (mixed DNA/RNA models) and compositional heterogeneity of base frequency (time-heterogeneous approach), is fundamental to robustly reconstruct phylogenies. Both tree topologies (time-heterogeneous and time-homogeneous) support the monophyly of Pterygota and the Chiasmomyaria scenario (Odonata basal) but the major clades, Hexapoda, Entognatha and Ectognatha are only recovered in the time-heterogeneous approach. However, the improved topology estimates by applying non-stationarity processes show still some weak supported ancient splits within Pterygota, especially within the basal Neoptera.

Focusing only on the interordinal relationships of insects, a new approach was conducted by expanding the data set of pterygotes (Simon & Hadrys 2010a, see 6.4).



Several new complete 28S rRNA sequences across a broad range of pterygote orders were added to the existing data set. Comprehensive phylogenetic analyses of the complete 28S rRNA gene on the sequence and secondary structure level were conducted to evaluate the potential of the gene to resolve deep ancient splits among insects. In addition phylogenetic analyses on a concatenated data set (18S+28S) were performed to evaluate the stability of the inferred phylogenetic relationships by the previous study (von Reumont et al. 2009, see 6.3).

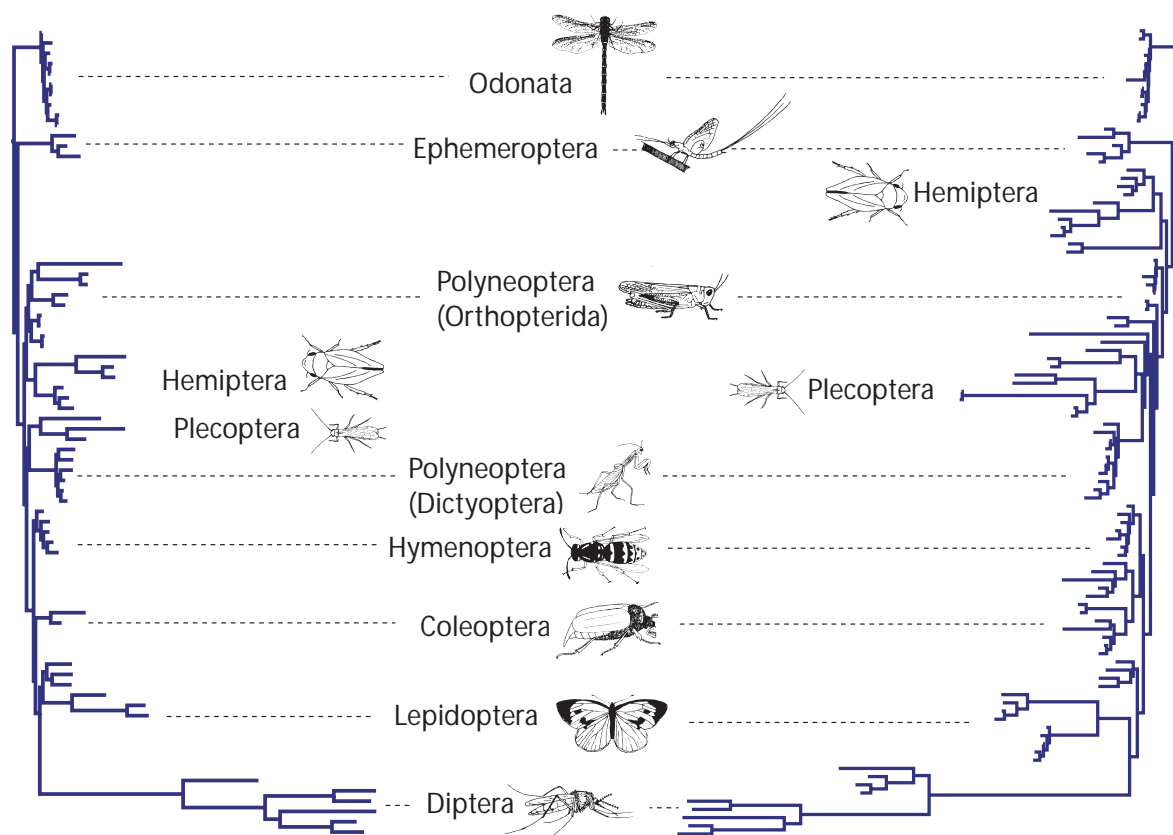


Figure 5: Comparison of the two time-heterogeneous consensus trees of the concatenated 18S+28S rRNA data set. Left: simplified topology based on the arthropod data set (von Reumont et al. 2009); right: simplified topology based on the hexapod data set (Simon & Hadrys 2010a). Complete consensus trees are shown in 6.3 and 6.4, respectively. Notice that the right tree has an improved pterygote taxon sampling. Pictures are modified after [112].

The analyses reveal that the 28S rRNA gene at both levels (sequence and secondary structure) do not harbor enough phylogenetic signal to recover ancient splits in Pterygota with artificial clades caused by signal erosion or occurrence of homoplasies. In contrast, analyses on the sequence level show that concatenation of the 28S and the 18S rRNA genes seems to compensate the signal erosion. Through comparison of both rRNA studies (von Reumont et al. 2009 and Simon & Hadrys 2010a) several incongruences among the

insect relationships are observed (Figure 5) which highlights the sensitivity of taxon sampling in insect phylogenetic studies based on rRNA data. Several phylogenetic relationships among the neopteran lineages are incongruent in the topologies, e.g. placement of Plecoptera, Hemiptera and Hymenoptera. In contrast, some inferred relationships remain stable: both studies show strong support for Odonata as the earliest branching pterygote lineage and Ephemeroptera as sister-group to the neopteran lineages (Chiaistomyaria), indicating a robust reconstructed phylogeny.

### 3.3 Phylogenomics

(Simon et al. 2009, Meusemann et al. 2010 and references therein)

Focusing on the two EST projects from the basal pterygote orders, Ephemeroptera and Odonata, a phylogenomic approach was applied to address the “Palaeoptera problem” and to evaluate phylogenetic informative proteins within the data sets (Simon et al. 2009, see 6.5). New developed bioinformatic tools were used (i) to identify putative core orthologs within the ESTs (HaMStR [59]) and (ii) to identify randomly similar positions within the alignment (ALIScore [61]). Two concatenated alignments were constructed to evaluate the support for each of the three hypotheses at the base of the pterygotes. The data sets differed in their represented taxa and genes, their content of missing data and consequently the overall number of characters (*maxspe*: 15 species, 125 genes, 31,643 amino acid positions and *maxgen*: 8 species, 150 genes, 42,541 amino acid positions).

To gain knowledge about the phylogenetic information within EST data sets the function of the represented genes were assessed and assembled in the four KOG (Eukaryotic Orthologous Groups) categories: (1) cellular processes and signaling, (2) information storage and processing, (3) metabolism and (4) poorly characterized (<ftp://ftp.ncbi.nih.gov/pub/COG/KOG/>).

Statistical and phylogenetic analyses of the data sets rejected the traditional Palaeoptera scenario as well as the Metapterygota scenarios and unambiguously support the Chiaistomyaria hypothesis (Odonata, Ephemeroptera+Neoptera), see Figure 6. Moreover, statistical tests of concatenated alignments based on their functional classification showed that proteins belonging to cellular processes and signaling seem to be more informative than those belonging to the other three categories. Interestingly the results are congruent with a phylogenetic study of the fungal kingdom [124]. The authors also identified proteins involved in cellular processes and signaling to be more informative

than others. Although more comparative data quality assessments need to be conducted, the two studies suggest that molecular markers (proteins) involved in cellular processes and signaling seem to harbor the most phylogenetic signal for resolving deep phylogenetic relationships.

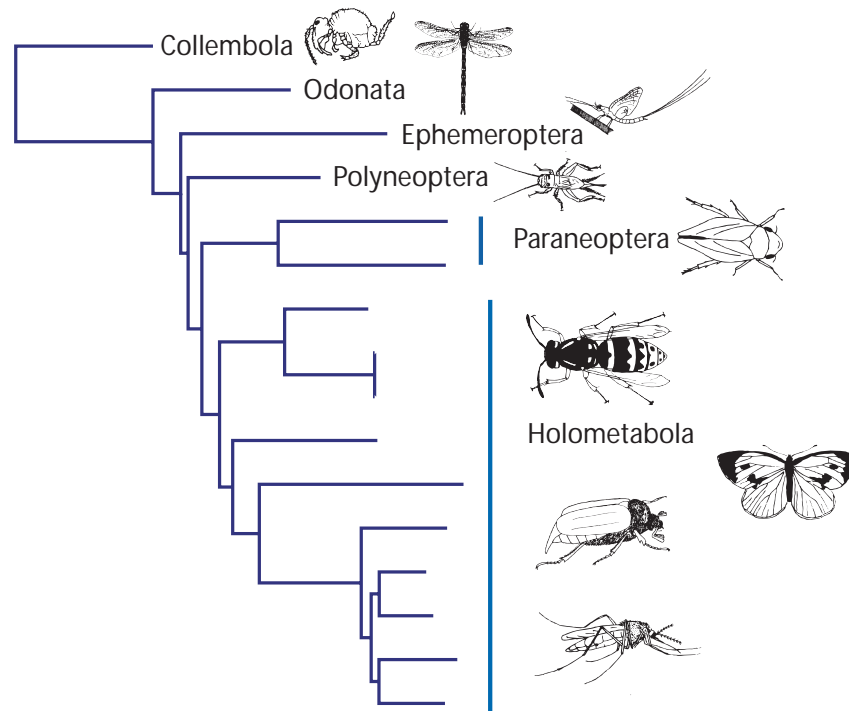


Figure 6: Simplified consensus topology of pterygote relationships based on the phylogenomic approach to infer basal pterygote divergence (Simon et al. 2009), see also 6.5. Pictures are modified after [112].

In addition a robust backbone tree of arthropods was reconstructed also by applying a phylogenomic approach (Meusemann et al. 2010, see 6.6). The initial data set was composed of 214 arthropods and 19 outgroup taxa derived from published genome- and EST-projects as well as from the collaborative projects within the DMPP-Arthropod group. Again gene orthology was assigned by applying the HaMStR approach [59] and randomly similar positions were identified by ALIScore [61]. In addition to increase the saturation of the data and to select an optimal subset MARE was applied (Misof et al., unpubl.). The inferred selected optimal subset (SOS) includes 117 taxa, 129 genes and 37,476 amino acid positions. Tree reconstruction was performed by maximum-likelihood analyses (ML) (RAxML 7.0.0 [125]) as well as Bayesian analyses (BI) (PhyloBayes 2.3c [126]). The optimized data set robustly resolves major arthropod relationships with strong support, in contrast to the tree based on the initial data set which is in many respects unresolved or shows low support values. However, the analyses on the optimized data set also show for

few deep splits among arthropods a remarkable sensitivity to methods of analyses, e.g. position of Ephemeroptera within Pterygota.

### 3.4 Mitogenomics

(Simon & Hadrys 2010b and references therein)

Two mitochondrial genomes of palaeopterous orders, of the of the dragonfly *Boyeria irene* and the mayfly *Baetis* sp., were isolated and characterized. Due to amplification problems of the control region and the amplification of nuclear pseudogenes (numts), the control region and flanking genes are missing for *Baetis* sp (see 6.7). The determined gene contents and orders of both taxa are identical to those of the common type of mitochondrial genomes observed in most insect orders. The two mt genomes are of crucial importance to re-evaluate the supported Metapterygota hypothesis of Zhang et al. [94] based on mitogenomic data.

The availability of complete mitochondrial genomes across hexapods (>170) makes them to a powerful source for comparative mitogenomics and phylogenetic studies on different taxonomic scales. However, no phylogenetic study has been performed on the complete data set. Most studies focused only on intraordinal relationships or interordinal relationships within the infraclasses and excluded several taxa/orders a priori, e.g. [26, 127-129]. The preliminary analyses performed here on the complete data set indicate the potential pitfalls in the analysis of mitogenomic data. The high evolutionary rate of mtDNA causing mutational saturation and homoplasy, compositional heterogeneity among lineages and heterogeneity of evolutionary rates among sites are all known to potentially bias inferred phylogenies from mitogenomic data. In addition, studies have shown the inconsistency between inferred relationships based on nuclear or mitochondrial datasets calling mitochondrial marker for resolving deep phylogeny into question [130, 131].

Maximum likelihood and Bayesian analyses were performed on a 31 taxa data set with the improved taxon sampling for the basal winged insect orders (four odonates and four mayflies). The phylogenetic analyses clearly demonstrate the potential main problem of mitogenomic data – the lack of phylogenetic signal for the deep nodes (Figure 7). Although monophyletic Pterygota was well recovered in the phylogenetic analyses, there was no significantly support for the earliest branching lineage of Pterygota. In general significant support was only recovered for the intraordinal relationships.

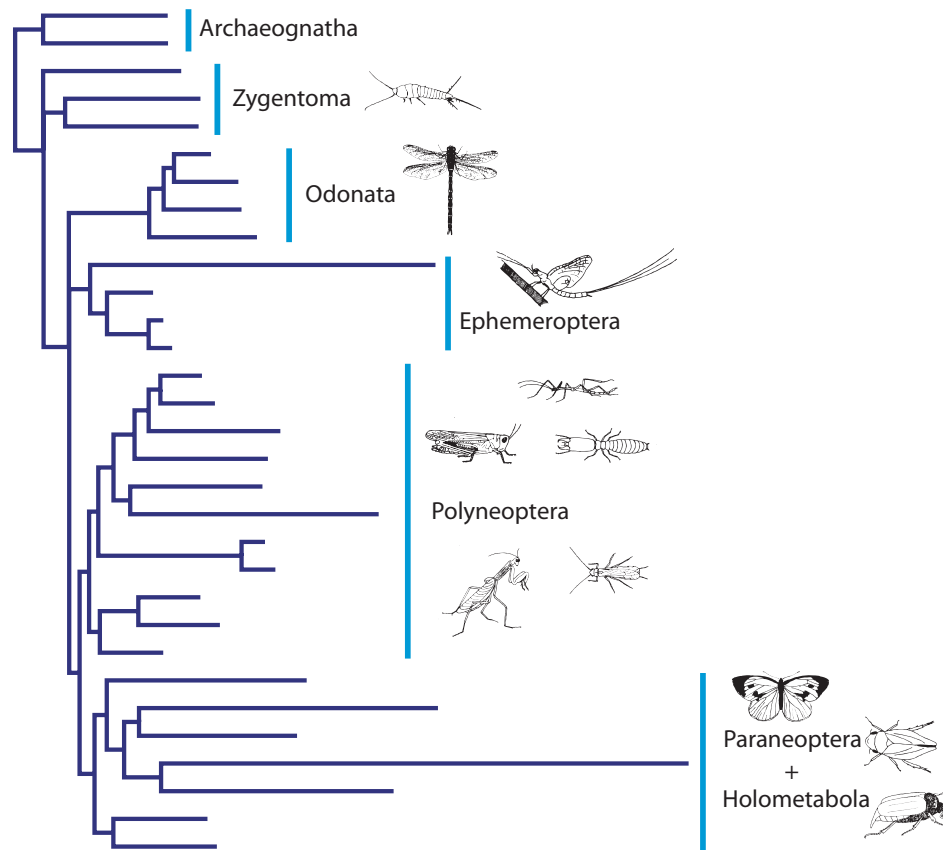


Figure 7: Simplified consensus topology of pterygote relationships based on the mitogenomic approach to resolve the Palaeoptera problem (Simon & Hadrys 2010b), see also 6.7. Pictures are modified after [112].

The usefulness of mitogenomic data for inferring relationships of highly divergent lineages is still controversial. In this context, Lin and Danforth [64] also argued that nuclear data sets should be preferred in deep insect molecular phylogenetic studies. In contrast, other studies have shown that mitochondrial genomes could be useful for deep phylogenetic relationships and are able to retrieve plausible phylogenetic relationships by applying appropriate substitution models [107, 132]. Still, these substitution models for mitochondrial sequence evolution might be first developed to robustly resolve the earliest branching lineages of Pterygota using mitogenomic data. Thus, based on the mitogenomic approach, the evolutionary history of the macro-evolutionary key transition – the origin of insect wings – remains an open question.

## 4. Conclusions

The thesis contributes several comprehensive molecular phylogenetic studies to resolve the insect evolutionary tree and in a grander scheme the arthropod evolutionary tree. Several molecular marker systems at different genetic levels from crucial insect orders were isolated and characterized providing the first step to get a deeper insight into the macro-evolutionary events among insects. The improved phylogenetic analyses demonstrated that data quality assessment prior to tree reconstruction yield better resolution in several places but the incongruence between the inferred topologies mirror the challenges in interpreting the strength and weakness of the underlying data and of the conducted analyses.

When handling molecular data we have always to consider that we are constructing a gene tree which reflects in the best-case scenario the species tree. However, because of the evolutionary pathway of a gene influenced by natural selection or genetic drift, gene trees and species trees are often inconsistent. Even if the “correct” gene tree is recovered, issues such as gene duplication and lineage sorting can yield to an erroneous species tree. Consequently, the evolutionary pathways of genes have a direct impact on the inferred topology and the phylogenetic signal is likely to be related to its evolutionary constraint. Through the application of different molecular marker systems, single nuclear protein coding genes, nuclear rRNA genes, ESTs and complete mitochondrial genomes, a variety of evolutionary processes and phylogenetic signal was covered. Therefore the different inferred pterygote relationships can be compared and allows the discrimination between a potential gene tree and species tree. In addition, the application of a multi-gene approach allowed us not only to increase the number of phylogenetic informative positions but also to identify potential genes which might evolve at the optimal rate and harbor the same evolutionary history along the branches as the speciation event under consideration.

However, the most convincing inferred relationship among pterygotes in this thesis is the well supported Chiasmomyaria hypothesis (Odonata, Ephemeroptera+Neoptera) based on independent data sets and applied analyses (rRNA genes and ESTs). The evolution of Pterygota and consequently insect wings is one of the most significant radiations among insects resulting into unimaginable species richness. Therefore determining the most basal pterygote order and hence uncovering the origin of insect wings and the evolution into a broad spectrum of sophisticated flight organs solved a major question in insect systematics. It is shown here that molecular systematic approaches are able to uncover this key innovation by closing the gap of genetic information, improved phylogenetic analyses and assessing the data quality. We have still to consider that phylogenetic analyses are not

protected against stochastic or systematic bias. Long-branch attraction coupled with taxon sampling, phylogenetic reconstruction methods and limited number of characters with base-composition bias are all potential pitfalls in molecular systematics. Unfortunately, no biological algorithm or evolutionary model currently covers the full complexity of biological history which can minimize the inconsistency of tree-reconstruction methods.

Nevertheless, the comprehensive analyses and inferred insect relationships of basal pterygotes presented in this thesis could provide a starting point/model – the odonates as the most basal pterygote order – for a variety of studies, such as the morphology of extant and fossil insects, physical modeling and evolutionary developmental research [133].

In sum, future phylogenetic analyses of insects should focus on (i) the basal neopteran divergences due to the still unresolved “comb” and incongruence between the inferred topologies, (ii) improvement of algorithm and evolutionary models for tree reconstruction methods and (iii) – most important – the data quality assessment in general.

## 5. References

1. Grimaldi DA, Engel MS: **Evolution of the Insects**. New York: Cambridge University Press; 2005.
2. Erwin TL: **Tropical forests: their richness in Coleoptera and other arthropod species**. *Coleopterists Bulletin* 1982, **36**(1):74-75.
3. Odegaard F: **How many species of arthropods? Erwin's estimate revised**. *Biological Journal of the Linnean Society* 2000, **71**(4):583-597.
4. Sherwood V: **Most Heat Tolerant**. Chapter 21 in *University of Florida Book of Insect Records, 2001* 1996, <http://entomology.ifas.ufl.edu/walker/ufbir/>.
5. Worland MR, Block W, Grubor-Lajsic G: **Survival of *Heleomyza borealis* (Diptera, Heleomyzidae) larvae down to - 60°C**. *Physiological Entomology* 2001, **25**(1):1-5.
6. Akers AA: **Adapted to Greatest Depth**. Chapter 19 in *University of Florida Book of Insect Records, 2001* 1996, <http://entomology.ifas.ufl.edu/walker/ufbir/>.
7. Hutchinson GE: **"Homage to Santa Rosalina or why are there so many kinds of animals?"** *Am Nat* 1959, **93**:145-159.
8. Evans HE: **Insect Biology: A Textbook of Entomology**. Addison Wesley Publishing Company; 1984.
9. Carpenter FM, Burnham L: **THE GEOLOGICAL RECORD OF INSECTS**. *Ann Rev Earth Planet Sci* 1985, **13**:297-314.
10. Carpenter FM: **Treatise on Invertebrate paleontology, Part R, Arthropoda 4, vols. 3 & 4: Superclass Hexapoda**. Lawrence: Geological Society of America & University of Kansas; 1992.
11. Gullan PJ, Cranston PS: **The Insects: An Outline of Entomology**, Third Edition edn: Blackwell Publishing; 2000.
12. Yang AS: **Modularity, evolvability, and adaptive radiations: a comparison of the hemi- and holometabolous insects**. *Evol Dev* 2001, **3**(2):59-72.
13. Mayhew PJ: **A tale of two analyses: estimating the consequences of shifts in hexapod diversification**. *Biological Journal of the Linnean Society* 2003, **80**:23-36.
14. Mayhew PJ: **Why are there so many insect species? Perspectives from fossils and phylogenies**. *Biol Rev Camb Philos Soc* 2007, **82**(3):425-454.
15. Mayhew PJ: **Shifts in hexapod diversification and what Haldane could have said**. *Proceedings of the Royal Society B: Biological Sciences* 2002, **269**:969 - 974.
16. Giribet G, Edgecombe GD, Wheeler WC: **Arthropod phylogeny based on eight molecular loci and morphology**. *Nature* 2001, **413**(6852):157-161.
17. Henning W: **Die Stammesgeschichte der Insekten**. Frankfurt am Main: Senckenbergische Naturforschende Gesellschaft; 1969.
18. Wheeler WC, Whiting MF, Wheeler QD, Carpenter JM: **The Phylogeny of the Extant Hexapod Orders**. *Cladistics* 2001, **17**:113-169.
19. Whiting MF, Carpenter JC, Wheeler QD, Wheeler WC: **The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology**. *Syst Biol* 1997, **46**(1):1-68.
20. Kristensen NP: **Phylogeny of extant hexapods**. In: *The Insects of Australia: A Textbook for Students and Research Workers*. Edited by Naumann ID, Lawrence, J.F., Nielsen, E.S., Spradberry, J.P., Taylor, R.W., Whitten, M.J., Littlejohn, M.J., vol. 125-140. Melbourne: CSIRO, Melbourne Univ. Press; 1991.
21. Boudreaux HB: **Arthropod phylogeny with special reference to insects**. New York: Wiley; 1979.



22. [http://www.kensmithpublishing.com/irish\\_mayflies.htm](http://www.kensmithpublishing.com/irish_mayflies.htm)  
[www.nicewallpapers.info/tr/insect-1.html](http://www.nicewallpapers.info/tr/insect-1.html)  
[www.birdcrossstitch.com/dragonflies/](http://www.birdcrossstitch.com/dragonflies/)  
<http://www.windsofkansas.com/insecta.html>  
<http://www.flickr.com/photos/bobisbob/1009502306/>
23. Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A: **Towards an 18S phylogeny of hexapods: accounting for group-specific character covariance in optimized mixed nucleotide/doublet models.** *Zoology (Jena)* 2007, **110**(5):409-429.
24. Kristensen NP: **Phylogeny of endopterygote insects, the most successful lineage of living organisms.** *Eur J Entomol* 1999, **96**:237-253.
25. Whiting MF: **Phylogeny of the holometabolous insect orders based on 18S ribosomal DNA: when bad things happen to good data.** *Exs* 2002(92):69-83.
26. Castro LR, Dowton M: **The position of the Hymenoptera within the Holometabola as inferred from the mitochondrial genome of *Perga condei* (Hymenoptera: Symphyta: Pergidae).** *Mol Phylogenet Evol* 2005, **34**(3):469-479.
27. Philippe H, Snell EA, Baptiste E, Lopez P, Holland PW, Casane D: **Phylogenomics of eukaryotes: impact of missing data on large alignments.** *Mol Biol Evol* 2004, **21**(9):1740-1752.
28. Savard J, Tautz D, Richards S, Weinstock GM, Gibbs RA, Werren JH, Tettelin H, Lercher MJ: **Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects.** *Genome Res* 2006, **16**(11):1334-1338.
29. Wiegmann BM, Trautwein MD, Kim JW, Cassel BK, Bertone MA, Winterton SL, Yeates DK: **Single-copy nuclear genes resolve the phylogeny of the holometabolous insects.** *BMC Biol* 2009, **7**:34.
30. Caterino MS, Cho S, Sperling FA: **The current state of insect molecular systematics: a thriving Tower of Babel.** *Annu Rev Entomol* 2000, **45**:1-54.
31. Gaunt MW, Miles MA: **An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks.** *Mol Biol Evol* 2002, **19**(5):748-761.
32. Engel MS, Grimaldi DA: **New light shed on the oldest insect.** *Nature* 2004, **427**(6975):627-630.
33. Whitfield JB, Kjer KM: **Ancient rapid radiations of insects: challenges for phylogenetic analysis.** *Annu Rev Entomol* 2008, **53**:449-472.
34. Kjer KM, Carle FL, Litman J, Ware J: **A molecular phylogeny of Hexapoda.** *Arthropod Systematics & Phylogeny* 2006, **64**:35-44.
35. Ogden TH, Whiting MF: **The problem with "the Paleoptera Problem:" sense and sensitivity.** *Cladistics* 2003, **19**:432-442.
36. Regier JC, Shultz JW: **Molecular phylogeny of arthropods and the significance of the Cambrian explosion for molecular systematics.** *Am Zool* 1998, **38**:918-928.
37. Hughes J, Longhorn SJ, Papadopoulou A, Theodorides K, de Riva A, Mejia-Chang M, Foster PG, Vogler AP: **Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles).** *Mol Biol Evol* 2006, **23**(2):268-278.
38. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD *et al*: **Broad phylogenomic sampling improves resolution of the animal tree of life.** *Nature* 2008, **452**(7188):745-749.

39. Roeding F, Hagner-Holler S, Ruhberg H, Ebersberger I, von Haeseler A, Kube M, Reinhardt R, Burmester T: **EST sequencing of Onychophora and phylogenomic analysis of Metazoa.** *Mol Phylogenet Evol* 2007, **45**(3):942-951.
40. Helmkampf M, Bruchhaus I, Hausdorf B: **Phylogenomic analyses of lophophorates (brachiopods, phoronids and bryozoans) confirm the Lophotrochozoa concept.** *Proc Biol Sci* 2008, **275**(1645):1927-1933.
41. Baurain D, Brinkmann H, Philippe H: **Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors?** *Mol Biol Evol* 2007, **24**(1):6-9.
42. Schierwater B, Eitel M, Jakob W, Osigus HJ, Hadrys H, Dellaporta SL, Kolokotronis SO, Desalle R: **Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis.** *PLoS Biol* 2009, **7**(1):e20.
43. Gadagkar SR, Rosenberg MS, Kumar S: **Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree.** *J Exp Zool B Mol Dev Evol* 2005, **304**(1):64-74.
44. Gatesy J, Baker RH, Hayashi C: **Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of Crocodylia.** *Syst Biol* 2004, **53**(2):342-355.
45. Wilkinson M, Cotton JA, Lapointe FJ, Pisani D: **Properties of supertree methods in the consensus setting.** *Syst Biol* 2007, **56**(2):330-337.
46. Gatesy J, Matthee C, DeSalle R, Hayashi C: **Resolution of a supertree/supermatrix paradox.** *Syst Biol* 2002, **51**(4):652-664.
47. Steel M, Rodrigo A: **Maximum likelihood supertrees.** *Syst Biol* 2008, **57**(2):243-250.
48. Cotton JA, Wilkinson M: **Supertrees join the mainstream of phylogenetics.** *Trends Ecol Evol* 2009, **24**(1):1-3.
49. de Queiroz A, Gatesy J: **The supermatrix approach to systematics.** *Trends Ecol Evol* 2007, **22**(1):34-41.
50. Felsenstein J: **Inferring phylogenies.** Sunderland (MA): Sinauer Associates, Inc.; 2004.
51. Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F: **Heterotachy and long-branch attraction in phylogenetics.** *BMC Evol Biol* 2005, **5**:50.
52. Lartillot N, Brinkmann H, Philippe H: **Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model.** *BMC Evol Biol* 2007, **7 Suppl 1**:S4.
53. Hartmann S, Vision TJ: **Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment?** *BMC Evol Biol* 2008, **8**:95.
54. Wiens JJ: **Missing data and the design of phylogenetic analyses.** *J Biomed Inform* 2006, **39**(1):34-42.
55. Wiens JJ: **Missing data, incomplete taxa, and phylogenetic accuracy.** *Syst Biol* 2003, **52**(4):528-538.
56. de la Torre-Barcelona JE, Kolokotronis SO, Lee EK, Stevenson DW, Brenner ED, Katari MS, Coruzzi GM, DeSalle R: **The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data.** *PLoS One* 2009, **4**(6):e5764.
57. Boero F, Schierwater B, Piraino S: **Cnidarian milestones in metazoan evolution.** *Integr Comp Biol* 2007, **47**:693-700.
58. Wägele JW, Mayer C: **Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects.** *BMC Evol Biol* 2007, **7**(1):147.

59. Ebersberger I, Strauss S, von Haeseler A: **HaMStR: profile hidden markov model based search for orthologs in ESTs.** *BMC Evol Biol* 2009, **9**:157.
60. Stocsits RR, Letsch H, Hertel J, Misof B, Stadler PF: **Accurate and efficient reconstruction of deep phylogenies from structured RNAs.** *Nucleic Acids Res* 2009, **37**(18):6184-6193.
61. Misof B, Misof K: **A Monte Carlo approach successfully identifies randomness of multiple sequence alignments.** *In press* 2008.
62. Whiting MF: **Mecoptera is paraphyletic: multiple genes and phylogeny of Mecoptera and Siphonaptera.** *Zoological Scripta* 2001, **31**:93-104.
63. Kjer KM: **Aligned 18S and insect phylogeny.** *Syst Biol* 2004, **53**(3):506-514.
64. Lin CP, Danforth BN: **How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets.** *Mol Phylogenet Evol* 2004, **30**(3):686-702.
65. Lewin B: **Genes.** New York: John Wiley & Sons; 1983.
66. Hughes J, Vogler AP: **The phylogeny of acorn weevils (genus Curculio) from mitochondrial and nuclear DNA sequences: the problem of incomplete data.** *Mol Phylogenet Evol* 2004, **32**(2):601-615.
67. Zakharov EV, Caterino MS, Sperling FA: **Molecular phylogeny, historical biogeography, and divergence time estimates for swallowtail butterflies of the genus Papilio (Lepidoptera: Papilionidae).** *Syst Biol* 2004, **53**(2):193-215.
68. Reed RD, Sperling FA: **Interaction of process partitions in phylogenetic analysis: an example from the swallowtail butterfly genus Papilio.** *Mol Biol Evol* 1999, **16**(2):286-297.
69. Mitchell A, Cho S, Regier JC, Mitter C, Poole RW, Matthews M: **Phylogenetic utility of elongation factor-1 alpha in noctuoidea (Insecta: Lepidoptera): the limits of synonymous substitution.** *Mol Biol Evol* 1997, **14**(4):381-390.
70. Cho S, Mitchell A, Regier JC, Mitter C, Poole RW, Friedlander TP, Zhao S: **A highly conserved nuclear gene for low-level phylogenetics: elongation factor-1 alpha recovers morphology-based tree for heliothine moths.** *Mol Biol Evol* 1995, **12**(4):650-656.
71. Belshaw R, Quicke DL: **A molecular phylogeny of the Aphidiinae (Hymenoptera: Braconidae).** *Mol Phylogenet Evol* 1997, **7**(3):281-293.
72. Regier JC, Shultz JW: **Molecular phylogeny of the major arthropod groups indicates polyphyly of crustaceans and a new hypothesis for the origin of hexapods.** *Mol Biol Evol* 1997, **14**(9):902-913.
73. Kamm K, Schierwater B, Jakob W, Dellaporta SL, Miller DJ: **Axial patterning and diversification in the cnidaria predate the Hox system.** *Curr Biol* 2006, **16**(9):920-926.
74. Brooke NM, Garcia-Fernandez J, Holland PW: **The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster.** *Nature* 1998, **392**(6679):920-922.
75. Galant R, Walsh CM, Carroll SB: **Hox repression of a target gene: extradenticle-independent, additive action through multiple monomer binding sites.** *Development* 2002, **129**(13):3115-3126.
76. Marden JH, Thompson JD: **Rowing locomotion by a stonefly that possesses the ancestral condition of co-occurring wings and abdominal gills.** *Biological Journal of the Linnean Society* 2003, **79**(2):341-349.
77. Wuyts J, De Rijk P, Van de Peer Y, Pison G, Rousseeuw P, De Wachter R: **Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA.** *Nucleic Acids Res* 2000, **28**(23):4698-4708.

78. Feng DF, Doolittle RF: **Progressive sequence alignment as a prerequisite to correct phylogenetic trees.** *J Mol Evol* 1987, **25**(4):351-360.
79. Olsen GJ, Woese CR: **Ribosomal RNA: a key to phylogeny.** *Faseb J* 1993, **7**(1):113-123.
80. Billoud B, Guerrucci MA, Masselot M, Deutsch JS: **Cirripede phylogeny using a novel approach: molecular morphometrics.** *Mol Biol Evol* 2000, **17**(10):1435-1445.
81. Letsch HO, Greve C, Kuck P, Fleck G, Stocsits RR, Misof B: **Simultaneous alignment and folding of 28S rRNA sequences uncovers phylogenetic signal in structure variation.** *Mol Phylogenet Evol* 2009, **53**(3):758-771.
82. Kjer KM: **Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs.** *Mol Phylogenet Evol* 1995, **4**(3):314-330.
83. Hickson RE, Simon C, Perrey SW: **The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence.** *Mol Biol Evol* 2000, **17**(4):530-539.
84. Hovmöller R, Pape T, Källersjö M: **The Palaeoptera Problem: Basal Pterygote Phylogeny Inferred from 18S and 28S rDNA Sequences.** *Cladistics* 2002, **18**:313-323.
85. Mallatt J, Giribet G: **Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch.** *Mol Phylogenet Evol* 2006, **40**(3):772-794.
86. Philippe H, Delsuc F, Brinkmann H, Lartillot N: **Phylogenomics.** *Annual Review of Ecology, Evolution and Systematics* 2005, **36**:541-562.
87. Delsuc F, Brinkmann H, Philippe H: **Phylogenomics and the reconstruction of the tree of life.** *Nat Rev Genet* 2005, **6**(5):361-375.
88. Philippe H, Derelle R, Lopez P, Pick K, Borchellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Queinnec E *et al*: **Phylogenomics revives traditional views on deep animal relationships.** *Curr Biol* 2009, **19**(8):706-712.
89. Roeding F, Hagner-Holler S, Ruhberg H, Ebersberger I, Arndt vH, Kube M, Reinhardt R, Burmester T: **EST sequencing of Onychophora and phylogenomic analysis of Metazoa.** *Mol Phylogenet Evol* 2007, **45**(3):942-951.
90. Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Baguna J, Bailly X, Jondelius U *et al*: **Assessing the root of bilaterian animals with scalable phylogenomic methods.** *Proc Biol Sci* 2009, **276**(1677):4261-4270.
91. Eisen JA, Fraser CM: **Phylogenomics: intersection of evolution and genomics.** *Science* 2003, **300**(5626):1706-1707.
92. DeSalle R: **Animal phylogenomics: multiple interspecific genome comparisons.** *Methods Enzymol* 2005, **395**:104-133.
93. Graur D, Li WH: **Fundamentals of molecular evolution.** Sunderland, Massachusetts: Sinauer Associates; 2000.
94. Zhang J, Zhou C, Gai Y, Song D, Zhou K: **The complete mitochondrial genome of Parafronurus youi (Insecta: Ephemeroptera) and phylogenetic position of the Ephemeroptera.** *Gene* 2008, **424**(1-2):18-24.
95. Yamauchi MM, Miya MU, Nishida M: **Use of a PCR-based approach for sequencing whole mitochondrial genomes of insects: two examples (cockroach and dragonfly) based on the method developed for decapod crustaceans.** *Insect Mol Biol* 2004, **13**(4):435-442.
96. Lee EM, Hong MY, Kim MI, Kim MJ, Park HC, Kim KY, Lee IH, Bae CH, Jin BR, Kim I: **The complete mitogenome sequences of the palaeopteran insects**

- Ephemera orientalis (Ephemeroptera: Ephemeridae) and Davidius lunatus (Odonata: Gomphidae).** *Genome* 2009, **52**(9):810-817.
97. Boore JL: **The use of genome-level characters for phylogenetic reconstruction.** *Trends Ecol Evol* 2006, **21**(8):439-446.
  98. Yang Z: **Among-site rate variation and its impact on phylogenetic analyses.** *Trends Ecol Evol (Amst)* 1996, **11**(9):367-372.
  99. Wiens JJ: **Can incomplete taxa rescue phylogenetic analyses from long-branch attraction?** *Syst Biol* 2005, **54**(5):731-742.
  100. Rodriguez-Ezpeleta N, Brinkmann H, Roure Be, Lartillot N, Lang BF, Philippe He: **Detecting and overcoming systematic errors in genome-scale phylogenies.** *Syst Biol* 2007, **56**(3):389-399.
  101. Brown JM, Lemmon AR: **The importance of data partitioning and the utility of bayes factors in bayesian phylogenetics.** *Syst Biol* 2007, **56**(4):643-655.
  102. **RNAseal** [<http://rnaseal.zfmk.de>]
  103. Gowri-Shankar V, Jow H: **PHASE: a software package for Phylogenetics And Sequence Evolution. 2.0.** In.: University of Manchester; 2006.
  104. Philippe He, Delsuc Fee, Brinkmann H, Lartillot N: **Phylogenomics.** *Annual Review of Ecology, Evolution, and Systematics* 2005, **36**(1):541-562.
  105. Misof B, Misof K: **A Monte Carlo Approach Successfully Identifies Randomness in Multiple Sequence Alignments : A More Objective Means of Data Exclusion.** *Systematic Biol* 2009, **58**(1):21-34.
  106. Talavera D, Hospital A, Orozco M, de la Cruz X: **A procedure for identifying homologous alternative splicing events.** *BMC Bioinformatics* 2007, **8**:260.
  107. Kjer KM, Honeycutt RL: **Site specific rates of mitochondrial genomes and the phylogeny of eutheria.** *BMC Evol Biol* 2007, **7**:8.
  108. Carapelli A, Lio P, Nardi F, van der Wath E, Frati F: **Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea.** *BMC Evol Biol* 2007, **7** Suppl 2:S8.
  109. Kjer KM, Carle FL, Litman J, Ware J: **A molecular phylogeny of Hexapoda.** *Arthropod Systematics & Phylogeny* 2006, **65**:35-44.
  110. Brady SG, Danforth BN: **Recent intron gain in elongation factor-1alpha of colletid bees (Hymenoptera: Colletidae).** *Mol Biol Evol* 2004, **21**(4):691-696.
  111. Goetze E: **Elongation factor 1-alpha in marine copepods (Calanoida: Eucalanidae): Phylogenetic utility and unique intron structure.** *Mol Phylogenet Evol* 2006, **40**(3):880-886.
  112. Sandhall A: **Insekten+Weichtiere: niedere Tiere und ihre Lebensräume - Gliedertiere, Würmer, Nesseltiere, Weichtiere, Einzeller.** München, Bern, Wien; 1974.
  113. McGinnis W, Krumlauf R: **Homeobox genes and axial patterning.** *Cell* 1992, **68**(2):283-302.
  114. Angelini DR, Kaufman TC: **Comparative developmental genetics and the evolution of arthropod body plans.** *Annu Rev Genet* 2005, **39**:95-119.
  115. Swalla BJ: **Building divergent body plans with similar genetic pathways.** *Heredity* 2006, **97**(3):235-243.
  116. Peel AD, Telford MJ, Akam M: **The evolution of hexapod engrailed-family genes: evidence for conservation and concerted evolution.** *Proc Biol Sci* 2006, **273**(1595):1733-1742.
  117. Ogishima S, Tanaka H: **Missing link in the evolution of Hox clusters.** *Gene* 2007, **387**(1-2):21-30.

118. Marden JH, O'Donnell BC, Thomas MA, Bye JY: **Surface-skimming stoneflies and mayflies: the taxonomic and mechanical diversity of two-dimensional aerodynamic locomotion.** *Physiol Biochem Zool* 2000, **73**(6):751-764.
119. Kjer KM, Baldrige GD, Fallon AM: **Mosquito large subunit ribosomal RNA: simultaneous alignment of primary and secondary structure.** *Biochim Biophys Acta* 1994, **1217**(2):147-155.
120. Schnare MN, Damberger SH, Gray MW, Gutell RR: **Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23 S-like) ribosomal RNA.** *J Mol Biol* 1996, **256**(4):701-719.
121. Amako D, Kwon OY, Ishikawa H: **Nucleotide sequence and presumed secondary structure of the 28S rRNA of pea aphid: implication for diversification of insect rRNA.** *J Mol Evol* 1996, **43**(5):469-475.
122. Gillespie J, Cannone J, Gutell R, Cognato A: **A secondary structural model of the 28S rRNA expansion segments D2 and D3 from rootworms and related leaf beetles (Coleoptera: Chrysomelidae; Galerucinae).** *Insect Mol Biol* 2004, **13**(5):495-518.
123. Gillespie JJ, Johnston JS, Cannone JJ, Gutell RR: **Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) rRNA genes of *Apis mellifera* (Insecta: Hymenoptera): structure, organization, and retrotransposable elements.** *Insect Mol Biol* 2006, **15**(5):657-686.
124. Kuramae EE, Robert V, Echavarri-Erasun C, Boekhout T: **Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom.** *BMC Evol Biol* 2007, **7**:134.
125. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**(21):2688-2690.
126. Lartillot N, Philippe H: **A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process.** *Mol Biol Evol* 2004, **21**(6):1095-1109.
127. Cameron SL, Barker SC, Whiting MF: **Mitochondrial genomics and the new insect order Mantophasmatodea.** *Mol Phylogenet Evol* 2006, **38**(1):274-279.
128. Cameron SL, Sullivan J, Song HJ, Miller KB, Whiting MF: **A mitochondrial genome phylogeny of the Neuropterida (lace-wings, alderflies and snakeflies) and their relationship to the other holometabolous insect orders.** *Zoologica Scripta* 2009, **38**(6):575-590.
129. Kim I, Cha SY, Yoon MH, Hwang JS, Lee SM, Sohn HD, Jin BR: **The complete nucleotide sequence and gene organization of the mitochondrial genome of the oriental mole cricket, *Gryllotalpa orientalis* (Orthoptera: Gryllotalpidae).** *Gene* 2005, **353**(2):155-168.
130. Timmermans MJ, Roelofs D, Marien J, van Straalen NM: **Revealing pancrustacean relationships: phylogenetic analysis of ribosomal protein genes places Collembola (springtails) in a monophyletic Hexapoda and reinforces the discrepancy between mitochondrial and nuclear DNA markers.** *BMC Evol Biol* 2008, **8**:83.
131. Wiens JJ, Kuczynski CA, Stephens PR: **Discordant mitochondrial and nuclear gene phylogenies in emydid turtles: implications for speciation and conservation.** *Biological Journal of the Linnean Society* 2010, **99**:445-461.
132. Jones M, Gantenbein B, Fet V, Blaxter M: **The effect of model choice on phylogenetic inference using mitochondrial sequence data: lessons from the scorpions.** *Mol Phylogenet Evol* 2007, **43**(2):583-595.

133. Yoshizawa K, Ninomiya T: **Homology of the wing base sclerites in Ephemeroptera (Insecta: Pterygota) - a reply to Willkommen and Hornschemeyer**. *Arthropod Struct Dev* 2007, **36**(3):277-279.

## 6. Publications and manuscripts upon which the thesis is based

- 6.1 Simon S, Schierwater B and H Hadrys (2010). On the value of Elongation factor-1 $\alpha$  for reconstructing Pterygote insect phylogeny, *Molecular Phylogenetics and Evolution* 54(2):651-656
- 6.2 Hadrys H, Simon S, Kaune B, Khadjeh S, Schmitt O, Schöner A and B Schierwater (2010). Isolation of Hox cluster genes from insects in development and evolution, *JEZ Part B: Molecular and Developmental Evolution*, in revision
- 6.3 von Reumont BM, Meusemann K, Szucsich NU, Dell'Ampio E, Gowri-Shankar V, Bartel D, Simon S, Letsch HO, Stocsits RR, Luan Y, Wägele JW, Pass G, Hadrys H and B Misof (2009). Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships, *BMC Evolutionary Biology* 9:119
- 6.4 Simon S & H Hadrys (2010). Comprehensive analysis of nuclear rRNA genes to infer Insect phylogeny, prepared for submission to *BMC Evolutionary Biology*
- 6.5 Simon S, Strauss S, von Haeseler A and H Hadrys (2009). A phylogenomic approach to resolve the basal pterygote divergence, *Molecular Biology and Evolution* 26(12):2719-2730
- 6.6 Meusemann K, von Reumont BM, Simon S, Roeding F, Kück P, Strauss S, Ebersberger I, Walz M, Pass G, Breuers S, Achter V, von Haeseler A, Burmester T, Hadrys H, Wägele JW and B Misof (2010). A phylogenomic approach to resolve the arthropod tree of life, *Molecular Biology and Evolution* (Epub 2010 June 9)
- 6.7 Simon S & H Hadrys (2010). The mitochondrial genome of two palaeopterous representatives: *Baetis* sp. (Ephemeroptera) and *Boyeria irene* (Odonata) – a mitogenomic approach to resolve the Palaeoptera problem, in preparation for *BMC Genome*



**On the value of Elongation factor-1 $\alpha$   
for reconstructing pterygote insect phylogeny**

**Sabrina Simon<sup>1,\*</sup>, Bernd Schierwater<sup>1,2</sup> and Heike Hadrys<sup>1,3</sup>**

<sup>1</sup> *ITZ, Ecology and Evolution, Stiftung Tierärztliche Hochschule Hannover, D-30559 Hannover, Germany*

<sup>2</sup> *American Museum of Natural History, New York City, NY 10024, USA*

<sup>3</sup> *Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06511, USA*

\*Corresponding author

This is the author's version of a work originally published by published by Elsevier in: *Molecular Phylogenetics and Evolution* 2010 Feb;54(2):651-656. Epub 2009 Oct 21; available under doi:10.1016/j.ympev.2009.09.029

**Abstract**

Pterygota are traditionally divided in two lineages, the “Palaeoptera” and Neoptera. Despite several efforts neither morphology nor molecular systematics have resolved the phylogeny of the pterygote insects. Too few markers have yet been identified for adequately tracking mesozoic-aged divergences. We tested the Elongation factor-1 $\alpha$  for its phylogenetic value in pterygote insect systematics. This highly conserved nuclear protein-coding gene has previously been reported to be useful in other groups for phylogenetic analyses at the intraordinal level as well as at the interordinal level. The analyses suggest that EF-1 $\alpha$  DNA sequences as well as intron positions provide informative markers for pterygote phylogenetics.

## 1. Introduction

The earliest divergence in the evolution of the winged insects formed two lineages, the basal "Palaeoptera" (insects without wing flexion muscles; Odonata and Ephemeroptera, plus several extinct orders) and the Neoptera (insects with wing flexion muscles). Despite the existence of numerous morphological characters and some large scale molecular data sets the relationships within the Pterygota remain ambiguous. The outstanding "Palaeoptera Problem", the unresolved "comb" among the basal Neoptera (including the positions of Dermaptera and Plecoptera) and the origin of the Holometabola (insects with complete metamorphosis) are some of the unresolved branching patterns (Kristensen, 1991; Whitfield and Kjer, 2008).

So far relatively few molecular markers have been employed to infer relationships between pterygote orders. The most commonly used molecular markers have been mitochondrial and nuclear rRNA gene fragments (Kjer, 2004; Ogden and Whiting, 2003). For studying deep splits in insects, nuclear ribosomal genes (18S and 28S) have been the most commonly used molecular markers (e.g. Kjer, 2004; Ogden and Whiting, 2003). These have led to concerns about different alignment methods and their effect on resulting phylogenies (Kjer, 2004; Terry and Whiting, 2005).

Consequently, the exploitation of new markers has become a crucial necessity for resolving pterygot insect phylogeny. Here nuclear coding genes may be promising candidates, since they evolve at a slower rate than mitochondrial coding genes and show little length variability. A few nuclear coding genes have become of wider use, including the Elongation factor-1 $\alpha$  (EF-1 $\alpha$  hereafter) which has proved useful in some terrestrial arthropods (reviewed in Caterino et al., 2000). Several studies have demonstrated that this marker is not only particularly useful for phylogenetic analyses among species groups and genera due to variation in silent nucleotide sites (Cho et al., 1995), but also for deeper divergences where amino acid replacements provide phylogenetic information (Regier and Shultz, 1997). EF-1 $\alpha$  was originally identified as a single-copy gene in insects. Later it was found that some insect orders possess multiple copies of EF-1 $\alpha$ , as these are Coleoptera, Hymenoptera, Diptera, Thysanoptera, Hemiptera-Coccoidae and members of the Neuropterida (reviewed in Djernaes and Damgaard, 2006). The paralogs show high nucleotide divergence between the two functional copies and are considered problematic in higher-level phylogenies of insects (Danforth and Ji, 1998; Hovemann et al., 1988; Jordal, 2002). The identification of a single ortholog in other Hemiptera (von Dohlen et al., 2002), Lepidoptera (Cho et al., 1995) and Odonata (Jordan et al., 2003) suggests independent

gene duplication in some orders, making EF-1 $\alpha$  problematic for generic level comparisons but not for phylogenetic studies targeting deeper divergence (Goetze, 2006; Lynch and Conery, 2000; Lynch and Conery, 2003). Compared to the often used ribosomal genes the EF-1 $\alpha$  gene is much less sensitive to alignment problems while its major disadvantage is that it is relatively short and best suited as a complement to other, ideally also high quality markers.

In insect systematics Elongation factor-1 $\alpha$  has been mostly used for studies within pterygote orders, for example in Hymenoptera systematics (Brady and Danforth, 2004), in Coleoptera systematics (Jordal, 2002), in Odonata systematics (Groeneveld et al., 2007), for resolving the phylogeny of the Neuropterida (Haring and Aspöck, 2004) and in a combined approach across insect orders (Kjer et al., 2006). Here we test previous expectations on the value of EF-1 $\alpha$  for insect systematics at the order level, and demonstrate its usefulness for inferring phylogenetic relationships among 20 pterygote insect orders. We focused on a 289bp coding region which is located within the suitable region suggested by Djernaes and Damgaard (2006). These authors argued to focus on a region between 493 and 1030 according to the mRNA transcript of the *Drosophila melanogaster* F1 copy based on their survey of intron positions and sequenced regions in Hexapoda. These authors focused on the exon-intron structure over a variety of hexapod orders and the phylogenetic value of these characters. Our purpose is to evaluate the phylogenetic value based on the nucleotide sequence of the recommend EF-1 $\alpha$  region including representatives of 20 pterygote orders. The study demonstrates the utility of the recommend EF-1 $\alpha$  region as molecular marker for phylogenetic analyses among pterygotes even though the amplified region is quite short.

## 2. Material and Methods

Species included in this study are given in Supplementary Material. DNA was extracted from ethanol (98%) preserved animals according to a modified standard protocol (Hadrys et al., 1993). For larger specimens a small tissue from a single leg or alternatively wing muscle from the mesothorax was dissected. For smaller specimens the entire thorax was used after removing of the gut. A 289-1039bp genomic fragment containing the 289bp coding region of EF-1 $\alpha$  was amplified via polymerase chain reaction (PCR) using the primers EF-7 [5'-AAC AAR ATG GAY TCN ACN GAR CCN CC-3'] and EF-9 [5'-CCN ACN GGB ACH GTT CCR ATA CC-3']. The amplification profile was as follows: initial

denaturation at 95°C for 5min followed by 45 cycles of 30sec at 94°C, 30sec at 60°C and 3min at 72°C. PCR products were checked and size determined by agarose gel electrophoresis. Amplification products from different species varied in size due to the presence of an intron of variable lengths (55bp to 546bp). PCR products were purified with MultiScreen PCR Plate (Millipore) following the manufacturer's instructions. Sequencing reactions were carried out in both directions using DYEnamic ET Dye Terminator Cycle Sequencing Kit (*Amersham Bioscience*). The sequencing reactions were purified with the Montage SEQ Kit (Millipore) following manufacturer's instructions and sequenced on a MegaBACE 1000 system (*Amersham Bioscience*).

Sequences were assembled and edited using SeqManII (vers. 5.00; DNASTAR, Inc.). Alignments were generated using ClustalX in MEGA4 (Tamura et al., 2007) and subsequently modified by analysing the intron/exon structure. Therefore the nucleic acid sequences were translated into the corresponding peptide sequences by Transeq (<http://www.ebi.ac.uk/emboss/transeq/>) for identification of coding regions. A minimum number of gaps in the non-coding region was inserted manually in order to produce reasonable conserved exon domains. Intron sequences were not included in the phylogenetic analyses, since attempts to align these turned out difficult due to high sequence divergence and the presence of long adenine and thymine repetitions. After removal of the intron the length of the aligned coding sequence was 289bp.

EF-1 $\alpha$  nucleotides were subjected to maximum likelihood (ML) and Bayesian analyses (BI). Prior to the analyses Modeltest 3.7 (Posada and Crandall, 1998) was used to determine an appropriate model of sequence evolution (best suited model according to the AIC criterion) for the ML analysis. Garli v.0.951 was used for performing the ML criterion. The assumed model of nucleotide substitution was the GTR (General Time Reversible) model with gamma distributed rate heterogeneity and an estimated proportion of invariable sites. Support values for ML trees were estimated with 1000 bootstrap replicates and 10 random sequence additions per replicate.

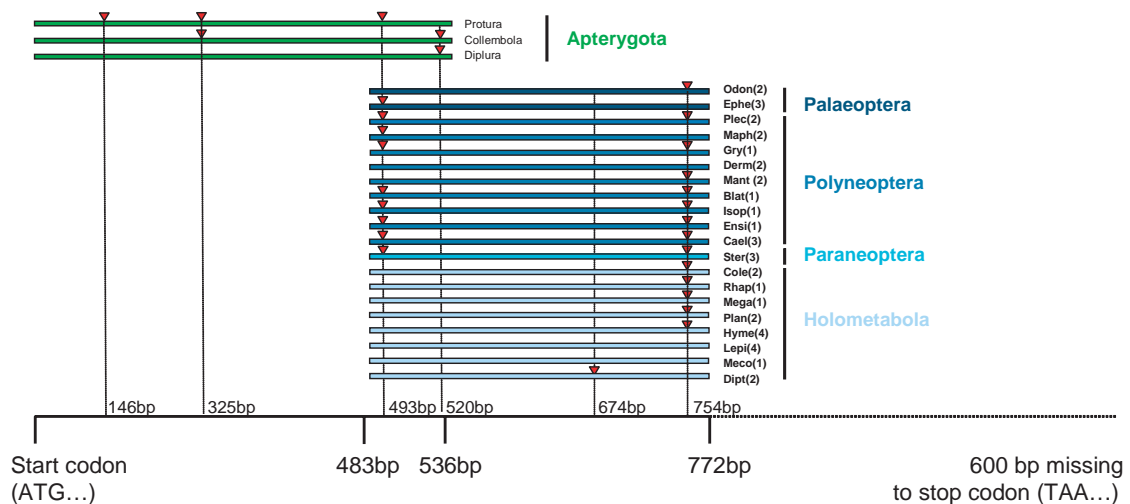
The BI analysis was conducted with MrBayes v.3.1.2 (Huelsenbeck and Ronquist, 2001) with a GTR+SSR model and a partitioned data set according to the codon positions, allowing independent parameter settings for each partition. The matrix was analyzed over 3.000.000 generations using four chains (one cold, three heated) and a sampling frequency of 100. Stationarity was evaluated graphically by plotting log-likelihood scores against generation and the first 7,500 trees were discarded as the "burn-in". The remaining trees were assembled into a topology. The sequence of the apterygote *Ctenolepisma lineata*

(Thysanura: GenBank accession no. **AF063405**) was used in all phylogenetic analyses as the outgroup.

### 3. Results

#### *Intron length and positions*

The isolated central region of the EF-1 $\alpha$  gene contained 289 nucleotides of coding sequence plus 55bp to 546bp of intron sequence across the 40 examined species. According to the mRNA transcript of the *Drosophila melanogaster* F1 copy (GenBank accession no. **X06869**) the amplified coding sequences start at position 483 (Fig.1). The comparative analysis of intron positions reveals some clade specific patterns. The longest observed intron contained 546 nucleotides (Plec2) and the shortest 55 nucleotides (Mega1), only the orders Lepidoptera, Dermaptera, Mecoptera, and one Diptera species (Dipt3) had no intron in the amplified region. All others had either one or two introns at two typical intron positions: (a) position 493 (shared by 9 orders) and (b) position 754 (shared by 14 orders). Only one dipteran species showed an intron in a different position (674; Fig.1).



**FIG. 1** – Amplified region of EF-1 $\alpha$  corresponding to the mRNA of *Drosophila melanogaster* F1 copy (GenBank accession no. **X06869**). Only coding regions are shown. In green: apterygote EF-1 $\alpha$  sequences with intron locations (Carapelli et al., 2000). Locations of red triangles indicate the intron positions of the different orders (number of species in parentheses). Most of the pterygote orders share two intron positions (position 493 and 754 respectively). Intron at position 493 does not occur in any holometabolous order. One intron position (674) occurs only in one Diptera species (Dipt2). For Lepidoptera, Dermaptera, Mecoptera and one Diptera species (Dipt3) no intron was found.

*Templates for spliceosomal introns*

From the total of 43 isolated introns, 41 introns had the initial and final two nucleotides which fit the canonical GT/AG template for spliceosomal introns. Only two species Blat1 (*Blaberus fusca*) and Hyme2 (*Bombus terrestris*) show the non-canonical GC/AG template for spliceosomal introns (data not shown). The non-canonical GC/AG template of Hyme2 was identified by comparison to other EF-1 $\alpha$  sequences of Hymenoptera (available in GenBank). Although four base pairs upstream a potential canonical GT template appears, this splicing site is not thought to be the template for spliceosomal introns because the peptide sequence would no longer be in frame with the normal EF-1 $\alpha$  protein. Splicing at the non-canonical GC/AG template creates the expected open reading frame. By comparison to other Hymenoptera sequences the non-canonical GC/AG template for spliceosomal introns appears as an apomorphy of the genus *Bombus* (bumble bees).

*Verification of EF-1 $\alpha$  amplicates*

Since EF-1 $\alpha$  is known to possess two paralogous copies amongst others in Hymenoptera, F1&F2 (Brady and Danforth, 2004; Danforth and Ji, 1998), Coleoptera, C1&C2 (Jordal, 2002) and Diptera, F1&F2 (Hovemann et al., 1988), we investigated whether the obtained sequences in this study were homologs of the original EF-1 $\alpha$  gene, which is the single EF-1 $\alpha$  copy reported from some Hemiptera (Heteroptera, Auchenorrhyncha, Sternorrhyncha (Cho et al., 1995; Normark, 1999; von Dohlen et al., 2002), Lepidoptera (Cho et al., 1995) and Odonata (Jordan et al., 2003)). Comparing our sequences from Hymenoptera, Coleoptera, and Diptera to known sequences from these orders suggests the exclusive presence of the ancestral ortholog copy (i.e. F2 in Hymenoptera, C1 in Coleoptera and F1 in Diptera). In all cases the genetic distance to the latter is lower than to the derived paralog copy (Table 1a-c). Furthermore, in Hymenoptera the new sequences share the intron position with the F2 copies of Hymenoptera while in F1 copies no intron appears at this position (data not shown). For the Neuropterida it was not possible to distinguish between different copies since in an earlier study it was not possible to found a clear distinction based on the coding sequence (Haring and Aspöck, 2004).

**Table 1** – (a) Genetic distances of the amplified EF-1 $\alpha$  sequences (289bp) to the two known copies F1 and F2 in Hymenoptera. GenBank accession nos.: *Apis melifera* F1 (**X52884**), *Andrena* sp. F2 (**AY230129**), *Apis melifera* F2 (**AF015267**), *Dufourea mulleri* F2 (**AF435383**), *Hesperapis larreae* F2 (**AY230131**). (b) Genetic distances of the amplified EF-1 $\alpha$  sequences (289bp) to the two known copies C1 and C2 in Coleoptera. GenBank accession nos.: *Coccotrypes impressus* C1 (**AF259874**), *Coccotrypes advena* C1 (**AF444076**), *Theoborus ricini* C1 (**AF186691**), *Xyleborus sphenos* C1 (**AF186692**), *Coccotrypes advena* C2 (**AF508928**), *Coccotrypes impressus* C2 (**AF508923**), *Theoborus ricini* C2 (**AF508927**), *Xyleborus sphenos* C2 (**AF508925**). (c) Genetic distances of the amplified EF-1 $\alpha$  sequences (289bp) to the two known copies F1 and F2 in Diptera. GenBank accession nos.: *Drosophila melanogaster* F1 (**X06869**), *Drosophila melanogaster* F2 (**X06870**).

(a)

	Hyme1	Hyme2	Hyme3	Hyme5	Apis F1	Andrena F2	Apis F2	Dufourea F2	Hesperapis F2
Hyme1									
Hyme2	0.16								
Hyme3	0.18	0.20							
Hyme5	0.21	0.24	0.20						
Apis F1	0.33	0.37	0.31	0.26					
Andrena F2	0.21	0.21	0.20	0.25	0.30				
Apis F2	0.17	0.11	0.22	0.24	0.36	0.20			
Dufourea F2	0.19	0.16	0.20	0.20	0.32	0.16	0.14		
Hesperapis F2	0.15	0.18	0.16	0.18	0.29	0.13	0.17	0.12	

(b)

	Cole1	Cole2	Cole5	Cocco. impr. C1	Cocco. adv. C1	Theoborus C1	Xyleborus C1	Cocco. adv. C2	Cocco. impr. C2	Theoborus C2	Xyleborus C2
Cole1											
Cole2	0.22										
Cole5	0.20	0.20									
Cocco. impr. C1	0.25	0.20	0.18								
Cocco. adv. C1	0.25	0.20	0.18	0.00							
Theoborus C1	0.23	0.22	0.24	0.05	0.05						
Xyleborus C1	0.22	0.26	0.16	0.05	0.05	0.07					
Cocco. adv. C2	0.27	0.34	0.32	0.34	0.34	0.29	0.34				
Cocco. impr. C2	0.22	0.34	0.32	0.34	0.34	0.29	0.34	0.05			
Theoborus C2	0.29	0.28	0.26	0.29	0.29	0.29	0.34	0.08	0.10		
Xyleborus C2	0.27	0.29	0.27	0.34	0.34	0.29	0.34	0.08	0.10	0.16	

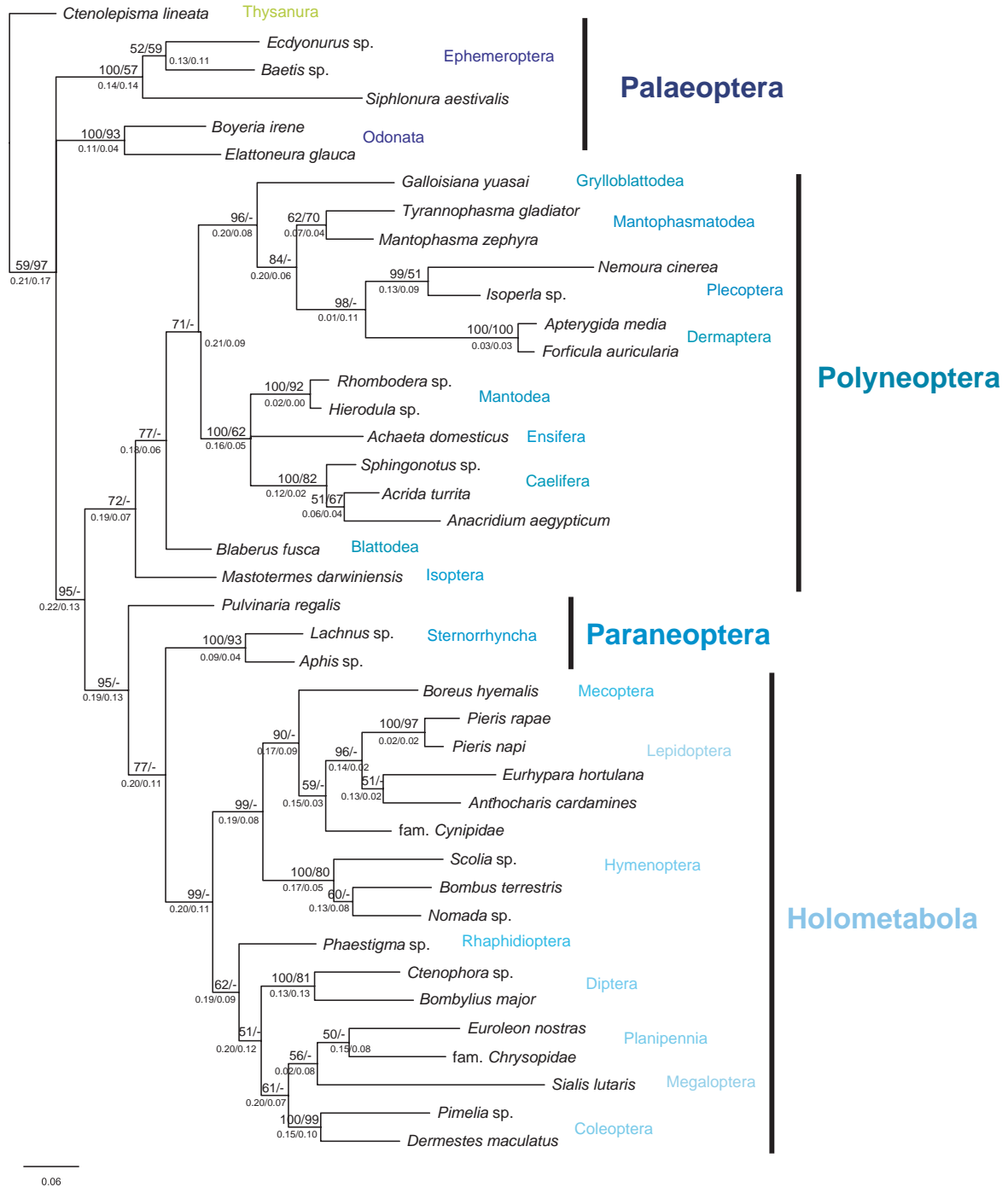


(c)

	Dipt2	Dipt3	Drosophila F1	Drosophila F2
Dipt2				
Dipt3	0.18			
Drosophila F1	0.31	0.27		
Drosophila F2	0.41	0.34	0.22	

#### *Phylogenetic information in Elongation factor-1 $\alpha$ exon sequence*

The variable - and not align able introns - were excluded prior to the phylogenetic analyses. The Bayesian analyses (independent parameter settings for codon positions) resulted a topology according to the infraclasses (Fig.2), except that Paraneoptera forms a paraphyletic group. Only the basal pterygote divergence could not be resolved and the majority rule consensus tree provides no support for any of the three hypotheses (Palaeoptera/Metapterygota/Chiasatomyria: Palaeoptera hypothesis 47% PP, Metapterygota hypothesis 45% PP, Chiasatomyria hypothesis no support). The maximum likelihood tree (gamma distributed rate heterogeneity) is generally identical to the Bayesian tree and resulted in a topology according to the infraclasses with the exception that the relationships between the orders vary. However, the bootstrap support values based on the ML analyses are only significant at the order-level (>50%). Based on the BI topology we also calculated mean pairwise distance between groups at both the nucleotide and amino acid level using MEGA4 (Tamura et al., 2007) (Fig.2). Although the sequence length is quite limited (289 bp, 96 aa) the pairwise distance values nonetheless increase with phylogenetically depth.



**FIG. 2** – Phylogeny of Pterygota based on phylogenetic analyses of EF-1 $\alpha$  nucleotide sequence. Consensus of topologies generated via MrBayes with 3,000,000 generations (first 7,500 trees were discarded as “burn-in”) using the GTR+SSR model. Numbers above nodes represent percentage of group inclusion among all topologies generated with MrBayes using the GTR+SSR model and bootstrap support percentages for 1000 replicates for maximum likelihood analyses respectively. Support values below 50% are not given. Numbers below nodes and separated by slashes represent mean pairwise divergence values between groups at nucleotide and amino acid level respectively. Insect orders in blue.

*Pattern of sequence variation*

Additional analyses also highlight the superioribility of EF-1 $\alpha$  over Histone H3. The sequence statistics summarized in Table 2 show that the codon positions have unequal base compositions but did not show a strong bias. The ML estimates of rate matrix and transition/transversion ratio reveal that base substitutions process is heterogeneous among the codon positions. In the third codon position transitions occur at higher frequency than transversions (three times more frequently in EF-1 $\alpha$  and about 4 times in Histone H3). The  $\alpha$  shape parameter estimated for the third position of EF-1 $\alpha$  is about 1 (0.99) indicating little or no among-site variation. In contrast, the third position of Histone H3 and the first and second position of Histone H3 and EF-1 $\alpha$  show  $\alpha$  values of 0.20-0.59 indicating considerable among-site variation. The proportion of invariable sites is decreasing from first to third position in EF-1 $\alpha$ . Histone H3 shows the highest amount of invariable sites in the second position (0.71).

As a further test for possible saturation of substitutions we plotted the observed number of transitions (ti) and transversions (tv) at each codon position against uncorrected (“p”) sequence divergence (Supplementary Material).

**Table 2** – Sequence statistics and substitution model parameters for Elongation factor-1 $\alpha$  and Histone H3 estimated on the ML topology. Base frequencies, rate matrix, percent invariant sites,  $\alpha$  shape parameter are estimated under the GTR+G+I model using PAUP\*. TS/TV ratio is estimated under the HKY85+G+I model.

	Character positions							
	EF-1 $\alpha$				Histone H3			
	all	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	all	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>
No. nt sites	288	96	96	96	327	109	109	109
Base frequencies								
A	0.33	0.29	0.37	0.36	0.20	0.18	0.30	0.16
C	0.19	0.19	0.20	0.17	0.31	0.31	0.27	0.31
G	0.22	0.38	0.16	0.21	0.24	0.32	0.21	0.26
T	0.25	0.15	0.27	0.26	0.25	0.19	0.22	0.27
Rate matrix R								
AC	2.93	0.73	4.45	0.40	3.68	78.94	>0.00	22.10
AG	6.50	1.56	1.53	19.99	27.80	2.41	0.82	1087.93
AT	3.46	4.25	1.10	9.43	6.13	52.70	>0.00	73.22
CG	2.19	0.21	10.14	1.45	0.52	>0.00	12.41	0.13
CT	11.37	3.11	1.30	34.04	8.10	72.11	>0.00	64.70
GT	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Percent invariant sites	0.36	0.42	0.13	0.06	0.43	0.50	0.71	0.00
$\alpha$ shape parameter	0.72	0.59	0.20	1.00	0.30	0.49	0.31	0.47
TS/TV ration	1.70	1.12	0.18	2.86	2.19	0.91	0.08	4.09

#### 4. Discussion

Whenever paralogs of a marker gene exist the use of this gene in phylogenetics may become problematic, at least at lower taxonomic levels. For the EF-1 $\alpha$  gene identification of paralogs in insects is comparatively easy, since in Hymenoptera, Coleoptera, Diptera the paralog copy has become notably diverged from the ancestral gene (Brady and Danforth, 2004; Danforth and Ji, 1998). In addition to sequence divergence the inclusion of an intron in the ancestral hymenopteran ortholog only (but not in the paralog copy) further eases ortholog identification. For Thysanoptera, Hemiptera-Coccoidae and members of the Neuropterida the distinction between the copies is more complicated since they show no differences in exon-intron structure nor in their coding sequence (Downie and Gullan, 2004; Haring and Aspöck, 2004; Morris et al., 2002). However, it seems that the newly designed primers show a high specificity to one orthologous EF-1 $\alpha$  copy only. We strongly believe that these primers enable us to isolate homologous EF-1 $\alpha$  copies for all of the 20 tested pterygote orders, thus allowing the use of this marker in any phylogenetic analyses of insects.

Two questions remain: (1) Are the paralog copies in Coleoptera, Hymenoptera and Diptera the results of independent duplication events or shared ancestry? (2) How ill-suited would be the use of EF-1 $\alpha$  paralogs instead of orthologs, if falsely identified? Danforth and Ji (1998) demonstrated that the F2 copy in Hymenoptera is not homologous to the F2 copy in Diptera and suggested independent gene duplication in both orders instead of an ancestral gene duplication event. In such cases of independent duplications of an essential gene that is continuously expressed throughout all life stages, one should expect that the original ortholog maintains its essential function while the extra copy may diverge and adopt new or complementing functions. As a consequence the ancestral ortholog will be widely conserved while the paralog copy will be subjected to an independent – and likely much more rapid – evolution. We suspect that this is the case for EF-1 $\alpha$  in pterygote insects. If, however, the obtained sequences of Hymenoptera, Coleoptera, Diptera and members of the Neuropterida were paralogous gene copies this would be problematic for phylogenetic analyses at lower taxonomic levels but at a lesser degree for phylogenetic analyses at the order level. Several studies have shown that paralogous gene copies are problematic in phylogenetic analyses at the species level but not when targeting deeper phylogenetic splits (Goetze, 2006; Lynch and Conery, 2000; Lynch and Conery, 2003). These studies together with our data presented here refute the concern in using EF-1 $\alpha$  as

molecular marker for higher-level phylogeny of insects (Danforth and Ji, 1998; Hovemann et al., 1988; Jordal, 2002).

Based on the presented analyses we confirm with an earlier study of Djaernes and Damgaard (2006) in using the recommend region of the EF-1 $\alpha$  gene. We are surprised that already the small exon fragment of EF-1 $\alpha$  analysed here provides a topology separating the infraclasses for the tested pterygote insect orders which is in agreement with evidence from other molecular markers and morphology. It is also noteworthy that the same phylogenetic analyses with the almost same taxon sampling using the Histone H3 gene resulted in no resolution at these deep nodes (order level; Supplementary Material) although it is used widely as molecular marker in insect phylogeny. The codon position comparison between EF-1 $\alpha$  and Histone H3 further supports this observation. Overall the rate matrix, TS/TV ratio,  $\alpha$  shape parameter and proportion of invariant sites varied in both genes across codon positions (Table 2). However, Histone H3 shows extreme conservation in the second position and substitutions occur mainly at silent sites (Table 2 and Supplementary Material). For EF-1 $\alpha$  the proportion of invariant sites at first and second codon position is smaller (Table 2) and substitutions occur more frequently than in Histone H3 (Supplementary Material). These analyses call for a replacement of Histone H3 and the use of EF-1 $\alpha$  (even though it is a nuclear protein coding gene) as an additional informative molecular marker in pterygote systematics, as previously suggested by Kjer et al. (2006). In addition, the overall signal quality is supported by the general increasing of the mean pairwise distance values at nucleotide as well as amino-acid level with phylogenetic depth, as expected for sequence data set well removed from saturation (Regier and Shultz, 1997).

The intron positions deserve separate attention, since it has been shown before that these “structural” characters may mirror relationships within insect orders and that recent intron insertions can characterize monophyletic groups (Brady and Danforth, 2004; Goetze, 2006). For example, the intron at position 754 according to the mRNA transcript is believed to be ancestral within arthropods, with some secondary intron losses in some groups (Goetze, 2006). The intron at position 493 is present in several pterygote orders and may be a pterygote synapomorphy. This intron appears in almost all basal, polyneopterous and paraneopterous orders but is absent in the derived holometabolous orders, indicating an intron loss in the evolution of the Holometabola. Coding these structural characters as intron presence/absence characters in phylogenetic analyses could provide additional “macromutational” information (Moulton and Wiegmann, 2004), as this has been shown

for example at an intraordinal level in bees (Brady and Danforth, 2004) and at an interordinal level in apterygote insects (Carapelli et al., 2000). However, Djernaes and Damgaard (2006) have shown that these structural characters can also cause erroneous grouping within pterygotes if the intron positions are used as major source of informative phylogenetic characters and suggest using them only as additional characters to a DNA sequence data set. Another interesting feature of the EF-1 $\alpha$  intron is the non-canonical GC/AG template for spliceosomal introns. This feature seems to be an apomorphy of the hymenoptera *Bombus* (bumble bees). The finding of a non-canonical GC/AG template for spliceosomal introns also in one species of Blattodea calls for additional data from this order before any conclusions can be drawn.

In sum, the presented data recommend Elongation factor-1 $\alpha$  as an informative nuclear marker to be added to molecular systematic studies of winged insects. The analyses show that the previously recommended region of the Elongation factor-1 $\alpha$  gene by Djernaes and Damgaard (2006) could be helpful to clarify some outstanding issues in pterygote systematics. We suggest continuing to focus on the genomic sequence of EF-1 $\alpha$  in combination with the intron structure analyses to gain additional information at both levels, gene structure and gene sequences. The possibility of amplifying paralogous copies exists but these can be identified in most cases and in any case should not affect phylogenetic analyses at higher taxonomic levels.

### **Acknowledgements**

We are grateful for the *Galloisiana yuasai* samples provided, collected and identified by Ryuichiro Machida, University of Tsukuba (Japan) in 2006. The Mantophasmatodea samples were kindly supplied by Reinhard Predel, University Jena and the *Mastotermes darwiniensis* samples by Horst Hertel, University Berlin. The remaining samples were mainly collected and identified by Albert Melber, Stiftung Tierärztliche Hochschule Hannover. This work was supported by a German Research Foundation (DFG) special priority program “Deep Metazoan Phylogeny” SP1174 grant given to H.H. (DFG HA 1947/5-1/2).

## References

- Brady S. G., Danforth B. N., 2004. Recent intron gain in elongation factor-1 $\alpha$  of colletid bees (Hymenoptera: Colletidae). *Mol Biol Evol* 21, 691-696.
- Carapelli A., Frati F., Nardi F., Dallai R., Simon C., 2000. Molecular phylogeny of the apterygotan insects based on nuclear and mitochondrial genes. *Pedobiologia* 44, 361-373.
- Caterino M. S., Cho S., Sperling F. A., 2000. The current state of insect molecular systematics: a thriving Tower of Babel. *Annu Rev Entomol* 45, 1-54.
- Cho S., Mitchell A., Regier J. C., et al., 1995. A highly conserved nuclear gene for low-level phylogenetics: elongation factor-1  $\alpha$  recovers morphology-based tree for heliothine moths. *Mol Biol Evol* 12, 650-656.
- Danforth B. N., Ji S., 1998. Elongation factor-1  $\alpha$  occurs as two copies in bees: implications for phylogenetic analysis of EF-1  $\alpha$  sequences in insects. *Mol Biol Evol* 15, 225-235.
- Djernaes M., Damgaard J., 2006. Exon-Intron Structure, Paralogy and Sequenced Regions of Elongation Factor-1  $\alpha$  in Hexapoda. *Arthropod Systematics & Phylogeny* 64 (1), 45-52.
- Downie D. A., Gullan P. J., 2004. Phylogenetic analysis of mealybugs (Hemiptera: Coccoidae: Pseudococcoidae) based on DNA sequences from three nuclear genes, and a review of the higher classification. *Systematic Entomology* 29, 238-259.
- Goetze E., 2006. Elongation factor 1- $\alpha$  in marine copepods (Calanoida: Eucalanidae): Phylogenetic utility and unique intron structure. *Mol Phylogenet Evol* 40, 880-886.
- Groeneveld L. F., Clausnitzer V., Hadrys H., 2007. Convergent evolution of gigantism in damselflies of Africa and South America? Evidence from nuclear and mitochondrial sequence data. *Mol Phylogenet Evol* 42, 339-346.
- Hadrys H., Schierwater B., Dellaporta S. L., DeSalle R., Buss L. W., 1993. Determination of paternity in dragonflies by Random Amplified Polymorphic DNA fingerprinting. *Mol Ecol* 2, 79-87.
- Haring E., Aspöck U., 2004. Phylogeny of the Neuropterida: a first molecular approach. *Systematic Entomology* 29, 415-430.
- Hovemann B., Richter S., Walldorf U., Cziepluch C., 1988. Two genes encode related cytoplasmic elongation factors 1  $\alpha$  (EF-1  $\alpha$ ) in *Drosophila melanogaster* with continuous and stage specific expression. *Nucleic Acids Res* 16, 3175-3194.
- Huelsenbeck J. P., Ronquist F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754-755.
- Jordal B. H., 2002. Elongation Factor 1  $\alpha$  resolves the monophyly of the haplodiploid ambrosia beetles Xyleborini (Coleoptera: Curculionidae). *Insect Mol Biol* 11, 453-465.
- Jordan S., Simon C., Polhemus D., 2003. Molecular systematics and adaptive radiation of Hawaii's endemic Damselfly genus *Megalagrion* (Odonata: Coenagrionidae). *Syst Biol* 52, 89-109.
- Kjer K. M., 2004. Aligned 18S and insect phylogeny. *Syst Biol* 53, 506-514.
- Kjer K. M., Carle F. L., Litman J., Ware J., 2006. A molecular phylogeny of Hexapoda. *Arthropod Systematics & Phylogeny* 65, 35-44.
- Kristensen N. P., 1991. Phylogeny of extant hexapods. In: *The Insects of Australia: A Textbook for Students and Research Workers* (ed. Naumann ID, Lawrence, J.F., Nielsen, E.S., Spradberry, J.P., Taylor, R.W., Whitten, M.J., Littlejohn, M.J.). CSIRO, Melbourne Univ. Press, Melbourne.
- Lynch M., Conery J. S., 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151-1155.

- Lynch M., Conery J. S., 2003. The evolutionary demography of duplicate genes. *J Struct Funct Genomics* 3, 35-44.
- Morris D. C., Schwarz M. P., Cooper S. J., Mound L. A., 2002. Phylogenetics of Australian Acacia thrips: the evolution of behaviour and ecology. *Mol Phylogenet Evol* 25, 278-292.
- Moulton J. K., Wiegmann B. M., 2004. Evolution and phylogenetic utility of CAD (rudimentary) among Mesozoic-aged Eremoneuran Diptera (Insecta). *Mol Phylogenet Evol* 31, 363-378.
- Normark B. B., 1999. Evolution in a putatively ancient asexual aphid lineage: recombination and rapid karyotype change. *Evolution* 53, 1458-1469.
- Ogden T. H., Whiting M. F., 2003. The problem with "the Paleoptera Problem:" sense and sensitivity. *Cladistics* 19, 432-442.
- Posada D., Crandall K. A., 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14(9), 817-818.
- Regier J. C., Shultz J. W., 1997. Molecular phylogeny of the major arthropod groups indicates polyphyly of crustaceans and a new hypothesis for the origin of hexapods. *Mol Biol Evol* 14, 902-913.
- Tamura K., Dudley J., Nei M., Kumar S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* 24, 1596-1599.
- Terry M. D., Whiting M. F., 2005. Comparison of two alignment techniques within a single complex data set: POY versus Clustal. *Cladistics* 21, 272-281.
- von Dohlen C. D., Kurosu U., Aoki S., 2002. Phylogenetics and evolution of the eastern Asian-eastern North American disjunct aphid tribe, Hormaphidini (Hemiptera: Aphididae). *Mol Phylogenet Evol* 23, 257-267.
- Whitfield J. B., Kjer K. M., 2008. Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annu Rev Entomol* 53, 449-472.



**Isolation of Hox cluster genes  
from insects in development and evolution**

**Heike Hadrys<sup>1,2\*</sup>, Sabrina Simon<sup>1</sup>, Barbara Kaune<sup>1</sup>, Sara Khadjeh<sup>1</sup>, Oliver Schmitt<sup>1</sup>,  
Anja Schöner<sup>1</sup> and Bernd Schierwater<sup>1,3</sup>**

<sup>1</sup> *ITZ, Division of Ecology and Evolution, Stiftung Tierärztliche Hochschule Hannover, Germany, 30559*

<sup>2</sup> *Dept. of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520*

<sup>3</sup> *American Museum of Natural History, New York City, NY 10024*

\*Corresponding author

This is the author's version of a work prepared for resubmission to *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*

**Abstract**

Among gene families it is the Hox genes and among metazoan animals it is the insects (Hexapoda) that have attracted particular attention for studying the evolution of development. Surprisingly, no Hox genes have been isolated from 26 out of 34 insect orders yet, and the existing sequences mainly derive from two orders only (61% from Hymenoptera and 22% from Diptera). We have isolated 37 partial homeobox sequences of Hox cluster genes (*lab*, *pb*, *Hox3*, *ftz*, *Antp*, *Scr*, *abd-a*, *Abd-B*, *Dfd*, and *Ubx*) from six insect orders, which are crucial to insect phylogenetics. These new gene sequences are a first step towards comparative Hox gene studies in insects in order to understand some key bauplan transitions. In addition, the deduced homeodomain sequences harbor phylogenetic information of potential relevance to insect systematics.

**Keywords**

Hox genes, insect specific Hox primers, degenerate Hox primers, Insects, Odonata, insect Evolution

## Introduction

*Antp*-class genes code for homeodomain-containing transcription factors that function in cell fate determination and embryonic development (e.g. McGinnis and Krumlauf, '92) . In Bilateria up to 100 *Antp*-class genes (including paralogs) can be divided into 30 gene families belonging to four major groups: HOX/PARAHOX genes (45 genes, four gene families), HOX-related genes (nine genes, five gene families), NK genes (16 genes, seven gene families), and NK-related genes (28 genes, 18 pseudogenes, 14 gene families). From the simplest Bilateria, the Platyhelmintha, 15 *Antp*-class genes are known and from the Arthropoda 37 (34 in Insecta). These genes have been of outstanding importance for metazoan radiation and provided deep insights into both, the phylogenetic patterns and the genetic mechanisms of animal bauplan development (e.g. Angelini and Kaufman, '05; Swalla, '06; Peel et al., '06; Ogishima and Tanaka, '07). Particularly Hox genes have attracted much attention since they define the identities of bauplan units (e.g. segments) along the anterior-posterior axis of the embryo (e.g. Sanchez-Herrero, '85; Angelini and Kaufman, '05). Hox genes have been known from all Bilateria and Hox-like genes also from diploblastic metazoans, including Placozoa and Cnidaria (e.g. Garcia-Fernandez, '05; Kamm and Schierwater, '06; Jakob and Schierwater, '07 for refs.).

Despite the importance of insects as the largest animal group on earth, and Hox genes as the most influential gene class in EvoDevo research, Hox genes have been isolated from only 8 out of some 34 insect orders yet. The full repertoire of Antennapedia genes has so far only been reported for *Folsomia candida*, *Tribolium castaneum* and *Drosophila melanogaster*. The majority of all sequences derive from two orders only, the Hymenoptera and the Diptera. In *Drosophila melanogaster* the Hox-Cluster is organized in two separate units: (a) the Antennapedia complex consisting of the Hox genes *labial* (*lab*), *proboscipedia* (*pb*), *Hox3* (*z2*, *zen*, *bcd*), *fushi tarazu* (*ftz*), *Deformed* (*Dfd*), *Sex combs reduced* (*Scr*) and *Antennapedia* (*Antp*), and (b) the Bithorax complex which includes *Ultrabithorax* (*Ubx*), *abdominal-A* (*abd-A*) and *Abdominal-B* (*abd-B*) (e.g. Scott, '93; Duncan, '87). This split is likely an aut-apomorphy of the Diptera since all of the above mentioned genes may be linked in a single cluster in other insects, e.g. Coleoptera (Beeman et al., '93).

It is highly unfortunate that very little is known about *Antp* genes in basal insects and that the origin and radiation of Hox genes in insects remains widely unresolved. Marden et al. ('00) highlight the crucial importance of isolating Hox genes particularly from basal Pterygota in order to reveal intermediate stages of evolution of appendages and

shed some light on the early evolution of flying insects. We here report on the successful isolation of 37 new homeobox fragments from six insect orders of crucial phylogenetic position, the apterygote Diplura and Archaeognatha, and the pterygote orders Ephemeroptera, Odonata, Plecoptera, and Dermaptera.

## Material and Methods

### Animal material and DNA extraction

*Campodea fragilis* (Diplura) and *Lepismachilis y-signata* (Archaeognatha) were kindly supplied by Karen Meusemann (ZFMK Bonn, Germany). *Sympetrum sanguineum*, *Ischnura elegans* (both Odonata) and *Baetis* sp. (Ephemeroptera) were collected at a small pond at our institute in Hannover, *Nemoura cinerea* (Plecoptera) was kindly supplied by National Museum Prague (Czechia) and *Forficula auricularia* (Dermaptera) was found in Hannover in a private garden. Tissue samples (legs of *S. sanguineum* or else whole animals) were preserved in ethanol (80%) and stored at 4°C. Whole genomic DNA was extracted according to Hadrys et al. ('92; '93).

### PCR amplification

Partial homeobox sequences of the genes *Deformed* (*Dfd*), *Sex combs reduced* (*Scr*), *Ultrabithorax* (*Ubx*) and *abdominal-A* (*abd-A*) were amplified by PCR with degenerate primers. We designed “insect specific” degenerate primers, which specifically amplify partial homeobox sequences of between 120 and 164bp of the target genes (Table 1). Alternatively, homeobox sequences were amplified by various combinations of four degenerated forward primers and five degenerated reverse primers reported in Cook et al. ('01).

*“Insect specific” degenerate Primer PCR:* Reactions were carried out in a total volume of 30 µl containing 40 pmol of each primer pair, 3.3 mmol of dNTP mix, and 1.5 U of Taq-Polymerase (Invitrogen). PCR started with an initial denaturation (93°C for 2 min) followed by 45 amplification cycles: denaturing at 92°C for 30 sec, annealing at 55 to 75°C (optimized for each primer pair and organism) for 35 sec, elongation at 72°C for 30 sec. All PCRs finished with a final elongation at 72°C for 5 min. PCR products were purified with Montage PCR Centrifugal Filter Devices (Millipore).

*Degenerate Primer PCR* (Cook et al., '01): The 50 µl reaction mix contained: 1x amplification buffer, 4 mM MgCl<sub>2</sub>, 0.2 mM dNTPs, 10 pM each primer and 0.04 U Taq

DNA polymerase (Bioline). The ramp up PCR started with an initial denaturation (95°C for 5 min) followed by 6 amplification cycles: denaturing at 94°C for 45 sec, annealing started at 48°C for 10 sec followed by a ramp to 56°C (0.1°C/sec) and a ramp to 72°C (0.2°C/sec), elongation at 72°C for 10 sec, and subsequent 30 amplification cycles: denaturing at 94°C for 30 sec, annealing started at 53°C for 10 sec followed by a ramp to 62°C (0.1°C/sec), elongation at 72°C for 30 sec and finished with a final elongation at 72°C for 5 min. PCR products of the expected length (~70 - ~100bp) were cut out of the gel and purified through ethanol precipitation.

**Table 1:** “Insect specific” degenerate primers for the amplification of *Dfd*, *Scr*, *Ubx*, and *abd-A* Hox genes.

<i>Name</i>	<i>Sequence (5' – 3')</i>	<i>AT (°C)</i>	<i>Fragment (bp)</i>
Dfd1fw Dfd1rev	CAAGCGGCAGCGGACNCSNTAYAC TCTTCCTCCGCACGTTCTTNGTRTTNGG	58 57	160
Scr1fw Scr1rev	GCAGCGGACCTCCTACACCMGNTAYCARAC TCATGGTGGCCATCTTGTGYTCYTTYTTCC	62 57	128
Ubx3fw Ubx3rev	GCCGGCAGACCTACACCMGNTAYCARAC CTCCTGCTCGTTCAGCTCYTTDATNGC	61 57	145
abdAfw abdArev	CGGCGGCGGGGNMGNCARAC GGGCCTGCTCGTTGATCTCYTTNACNGC	59 60	164

Given are the primer sequences (forward = fw, reverse = rev), optimal annealing temperatures (AT) and expected fragment length of PCR products.

### Cloning and sequencing

The purified products were A-tailed and inserted into the pGEM-T plasmid vector (Promega) and cloned into *E. coli* (Invitrogen) following the manufacturer’s instructions. Clones were sequenced in both directions on an ABI PRISM 310 Genetic Analyzer (Applied Biosystems) using BigDye® Terminator Cycle Sequencing Kit (v.1.1, Applied Biosystems). Sequences were analyzed and aligned using SeqMan II 5.03 (DNASTar, Lasergene) and ClustalW (Higgins, '89).

### Phylogenetic analyses

Due to the known duplication of Hox3 in metazoan evolution and the resulting paralogy problems, this homeobox fragment was excluded from phylogenetic analyses (cf. Panfilio and Akam, '07).

To evaluate the phylogenetic signal of the homeoboxes we concatenated nine homeoboxes (*lab*, *pb*, *Dfd*, *Scr*, *ftz*, *Antp*, *Ubx*, *abd-A*, *Abd-B*; resulting sequence length: 1620 nucleotides) from several insect orders (GenBank accession numbers are given in Supplement Table S1) and performed Maximum Likelihood (ML) and Bayesian Inference (BI) analyses. ML analyses (100 bootstrap replicates) were conducted using the Dnaml program of PHYLIP 3.68 (Felsenstein, '08) based on the multiple alignment of the nucleotide sequences. The transition/transversion ratio was set at 2.0, and 8 replicates of random-order taxon addition were performed by using the Jumble option. Bayesian analysis (GTR+G+I) was conducted using MrBayes v.3.2 (Ronquist and Huelsenbeck, '03). Metropolis-coupled Markov chain Monte Carlo (MCMCMC) sampling was carried out with one cold and three heated chains starting from random starting trees and the program default prior probabilities on model parameters. Analyses were run for 5,000,000 generations and samples of the Markov chain were taken every 100 generations. Bayesian posterior probabilities were obtained from the majority rule consensus of the tree sampled after the initial burn-in period (1,000).

### Results and Discussion

In this study we have isolated the first homeobox sequences of Hox cluster genes from six insect orders: Diplura (*lab*, *Dfd*, *Scr*, *Antp*, *ftz*, *abd-A*, *Abd-B*), Archaeognatha (*Dfd*, *Scr*, *Antp*, *Ubx*, *abd-A*, *Abd-B*), Ephemeroptera (*Dfd*, *Scr*, *Antp*, *Ubx*, *abd-A*, *Abd-B*), Odonata (*lab*, *pb*, *Hox3*, *Dfd*, *Scr*, *Antp*, *Ubx*, *abd-A*, *Abd-B*), Plecoptera (*Dfd*, *Scr*, *Antp*, *ftz*, *Ubx*, *Abd-B*), and Dermaptera (*Dfd*, *Scr*) (amino acid alignments are shown in Table 2).

**Table 2: Alignment of 37 new hexapod Hox gene homeodomains.**

The newly isolated sequences of *lab*, *pb*, *Dfd*, *Scr*, *ftz*, *Antp*, *Ubx*, *abd-A* and *abd-B* from the thysanuran *Campodea fragilis* (C.f.), the archaeognath *Lepismachilis y-signata* (L.y.), the odonates *Ischnura elegans* (I.e.) and *Sympetrum sanguineum* (S.s.), the ephemeropteran *Baetis* sp. (B.sp.), the plecopteran *Nemoura cinerea* (N.c.) and the dermapteran *Forficula auricularia* (F.a.) are aligned to their *Drosophila melanogaster* (D.m.) homolog. Dots indicate identical position.

Gene	Alignment					
	1	10	20	30	40	50
<b>lab</b>	.....	.....	.....	.....	.....	.....
D.m.	NNSGRTNFTN	KQLTELEKEF	HFNRYLTRAR	RIEIANLTLQL	NETQVKIWFQ	NRRMKQKKRV
C.f.			...K.....	.....SA...	.....	
S.s.			...K.....	..D..SA...	.....	
<b>pb</b>	.....	.....	.....	.....	.....	.....
D.m.	PRRLRTAYTN	TQLLELEKEF	HFNKYLCPRR	RIEIAASLDL	TERQVKVWFQ	NRRMKHKRQT
S.s.						
<b>Dfd</b>	.....	.....	.....	.....	.....	.....
D.m.	PKRQRTAYTR	HQILELEKEF	HYNRYLTRRR	RIEIAHTLVL	SERQIKIWFQ	NRRMKWKKDN
C.f.						
L.y.						
I.e.			.F.....	.....S.C.		
S.s.			.F.....	.....S.C.		
B.sp.			.F.....	.....S.N.		
F.a.						
<b>Scr</b>	.....	.....	.....	.....	.....	.....
D.m.	TKRQRTSYTR	YQTELELEKEF	HFNRYLTRRR	RIEIAHALCL	TERQIKIWFQ	NRRMKLKEH
C.f.						
L.y.						
S.s.						
I.e.						.....W.....
B.sp.						
F.a.						
<b>ftz</b>	.....	.....	.....	.....	.....	.....
D.m.	SKRTRQTYTR	YQTELELEKEF	HFNRYLTRRR	RIDIANALSL	SERQIKIWFQ	NRRMKSKKDR
C.f.			.....L.....	..E..HS.G.	T....	
N.c.			...K.L.....	..E..HS.T.	T....	
<b>Antp</b>	.....	.....	.....	.....	.....	.....
D.m.	RKRGRQTYTR	YQTELELEKEF	HFNRYLTRRR	RIEIAHALCL	TERQIKIWFQ	NRRMKWKKEN
C.f.						
L.y.						
S.s.						
I.e.						
B.sp.						
F.a.						
<b>Ubx</b>	.....	.....	.....	.....	.....	.....
D.m.	RRRGRQTYTR	YQTELELEKEF	HTNHYLTRRR	RIEIAHALCL	TERQIKIWFQ	NRRMKLKKEI
L.y.						
S.s.						
I.e.						
B.sp.						
<b>abd A</b>	.....	.....	.....	.....	.....	.....
D.m.	RRRGRQTYTR	FQTELELEKEF	HFNHYLTRRR	RIEIAHALCL	TERQIKIWFQ	NRRMKLKEH
C.f.						
L.y.						
B.sp.						
I.e.						.....NNS
S.s.						

Gene	Alignment					
	1	10	20	30	40	50
<b>Abd B</b>	.....	.....	.....	.....	.....	.....
D.m.	VRKKRKPYSK	FQTLELEKEF	LFNAYVSKQK	RWELARNLQL	TERQVKIWFQ	NRRMKNNKNS
C.f.	..	.....	.....	.....N.	.....	.....
L.y.	..	.....	.....	.....N.	.....	.....
I.e.	....S.....	.....	.....	.....N.	.....	..
B.sp.	.....	.....	.....	.....N.	.....	.....
N.c.	..	.....	.....	.....N.	.....	.....

These 37 new sequences fill in crucial gaps both at the base of insects as well as at the base of Pterygota (Table 3). As seen in the Table, we raised the number of insect orders with reported Hox cluster gene sequences from 8 to 14 and the number of known gene sequences in the matrix from 67 to 101. In these numbers we include sequences from the 8 Hox genes (*lab*, *pb*, *Dfd*, *Scr*, *Antp*, *Ubx*, *abd-A*, *abd-B*) as well as from the two homeotic genes, *Hox3* (*bicoid*) and *ftz*, which are integrated in the insect Hox cluster (or clusters in the case of Diptera).



**Table 3: Number of Hox genes known from the different insect orders.**

The full complement of Hox cluster genes has so far been known from Collembola, Diptera, and Coleoptera only. Partial information now includes 15 insect orders, and no information is available from at least 19 orders. Here = this study.

Order (Infraclass)	<i>lab</i>	<i>pb</i>	<i>Hox</i> 3	<i>Dfd</i>	<i>Scr</i> *	<i>ftz</i>	<i>Antp</i> *	<i>Ubx</i>	<i>abd-A</i>	<i>Abd-B</i>
<b>Diplura</b>	here	-	-	here	here	here	here	-	here	here
Collembola	1	1	1	1	1	1	1	1	1	1
Protura	-	-	-	-	-	-	-	-	-	-
<b>Archaeognatha</b>	-	-	-	here	here	-	here	here	here	here
Thysanura	1	1	-	1	1	1	1	-	1	1
<b>Ephemeroptera</b>	-	-	-	here	here	-	here	here	here	here
<b>Odonata</b>	here	here	here	here	here	-	here	here	here	here
<b>Plecoptera</b>	-	-	-	here	here	here	here	here	-	here
<b>Dermaptera</b>	-	-	-	here	here	-	here	-	-	-
Orthoptera	1	-	-	1	2	-	2	2	2	1
Hemiptera	1	1	-	1	1	-	1	1	1	1
Hymenoptera	2	2	-	1	2	-	2	37	>100	2
Diptera	8	4	1	3	3	5	5	7	8	6
Coleoptera	1	1	1	1	1	1	1	1	1	1
Lepidoptera	-	-	-	2	2	-	3	4	3	1
Embioptera, Notoptera, Mantodea, Mantophasmatodea, Blattodea, Isoptera, Phasmatodea, Zoraptera, Psocoptera, Phthiraptera, Thysanoptera, Hemiptera (Hemimetabola)  Megaloptera, Raphidioptera, Neuroptera, Trichoptera, Mecoptera, Siphonaptera, Strepsiptera (Holometabola)	No data available									

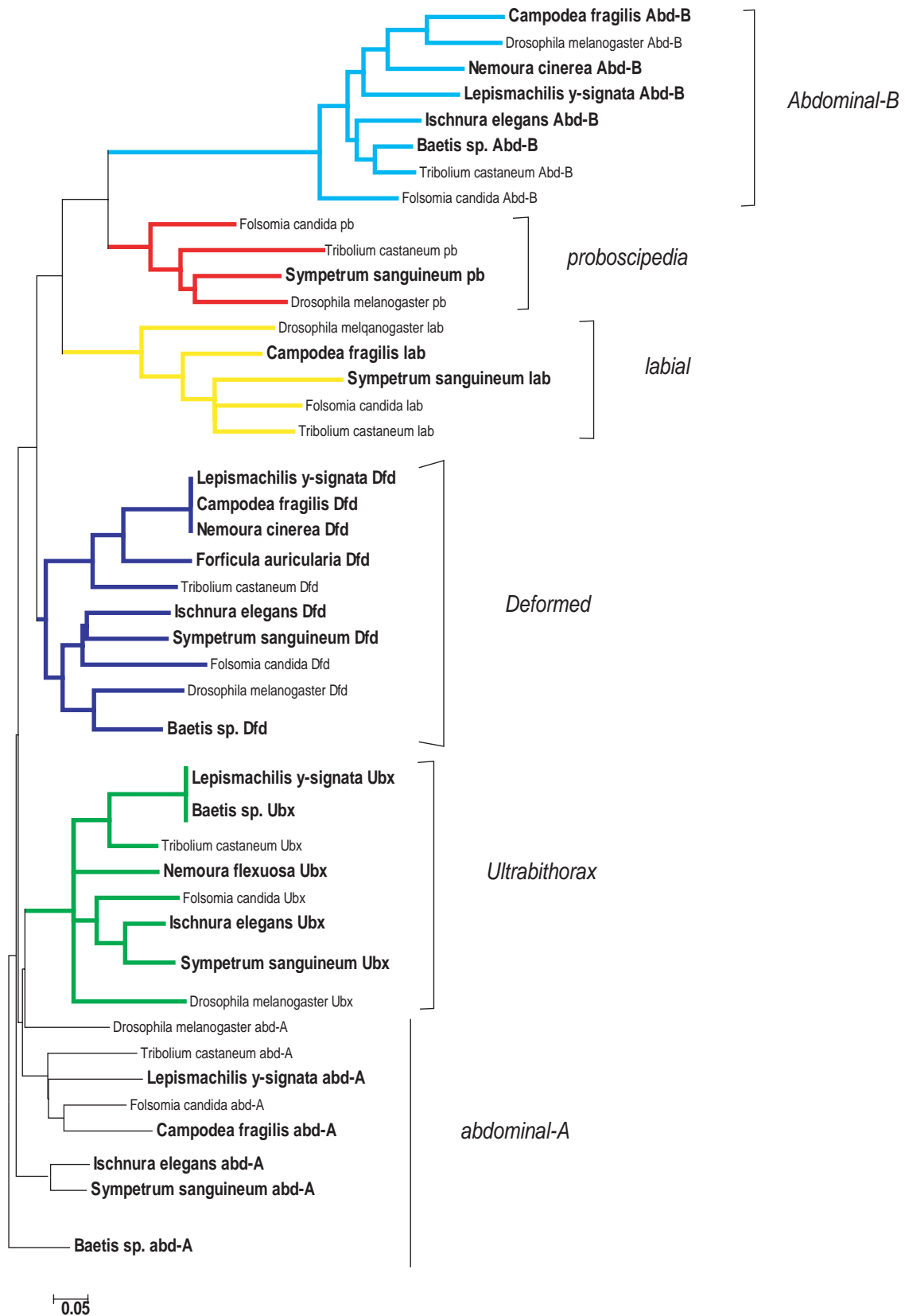
\* an unambiguous distinction between *Scr* and *Antp* based on short sequence fragment is not possible (see text).

Homolog identification of the isolated Hox genes is widely, but not completely non-problematic. All assignments shown in Table 2 are the immediate assignments according to BLAST searches. Phylogenetic analysis (Neighbor-Joining, NJ) of published homeobox sequences shows that the homeobox sequences for *lab*, *pb*, *Dfd*, *Ubx*, *abd-A*, *Abd-B* group into the expected clades whereas the complete homeodomain sequences for *Scr*, *ftz*, *Antp* do not group together in a simple distance analysis (Supplemental material Fig. S1). Therefore we performed another phylogenetic analysis (NJ) with only the six potentially unambiguous new homeobox sequences (*lab*, *pb*, *Dfd*, *Ubx*, *abd-A*, *Abd-B*). All new fragments group into the expected clades of homologs from other insects and thus confirm the results of NCBI Blast (Fig. 1). At present we cannot unambiguously distinguish between the *Scr* versus *Antp* homeobox fragments in Diplura, Archaeognatha, Ephemeroptera, Odonata, Plecoptera and Dermaptera. The isolated homeobox fragments differ for both genes and BLAST searches reveal assignments. However, no amino acid substitutions are found in this short fragment spanning homeodomain positions 20 to 45 (Table 4). For these gene fragments more sequence information is required to distinguish between the two alternatives since amino acid substitutions have been known to occur at positions 1, 4, 6, 7 and 60 only (Table 4). We believe that we have amplified both genes (different homeobox sequences) but we are reluctant to suggest an assignment to the *Scr* or *Antp* gene family in the absence of unambiguous differences in the homeodomain.

**Table 4: Alignment of *Scr* and *Antp* homeodomain fragments.**

From *Tribolium castaneum* (AF228509, AF227628); *Drosophila melanogaster* (M20705, X05228) and from insect species. For all new species the short *Scr* and *Antp* fragments differ in their homeobox sequence but are identical at the amino acid level. Amino acid substitutions between *Scr* and *Antp* are indicated with green and yellow, respectively.

	1	10	20	30	40	50	60
<i>Tribolium castaneum Scr</i>	T	K	R	G	R	T	S
<i>Tribolium castaneum Antp</i>	R	K	R	G	R	T	S
<i>Drosophila melanogaster Scr</i>	T	K	R	G	R	T	S
<i>Drosophila melanogaster Antp</i>	R	K	R	G	R	T	S
<i>Baetis sp. Scr/Antp</i>	-----	-----	-----	-----	-----	-----	-----
<i>Nemoura cinerea Scr/Antp</i>	-----	-----	-----	-----	-----	-----	-----
<i>Sympetrum sanguineum Scr/Antp</i>	-----	-----	-----	-----	-----	-----	-----
<i>Campodea fragilis Scr/Antp</i>	-----	-----	-----	-----	-----	-----	-----
<i>Lepismachilis y-signata Scr/Antp</i>	-----	-----	-----	-----	-----	-----	-----
<i>Forficula auricularia Scr/Antp</i>	-----	-----	-----	-----	-----	-----	-----



**Figure 1:** Neighbor-Joining tree of the new Hox gene sequences (*lab*, *pb*, *Dfd*, *Ubx*, *abd-A*, *Abd-B*) and known orthologs from other insects (GenBank accession numbers: *Folsomia candida* AF361326, AF361327, AF361329, AF361333, AF361334, AF361335; *Drosophila melanogaster*; NM\_057265, X63728, X05136, X76210, X54453, X16134; *Tribolium castaneum* AF231104, AF187068, U81039, AF146649, AF017415, AF227923). Sequences from this study are in bold. Note that all new sequences group to expected homologs.

The only Hox gene sequences previously isolated from Apterygota were from two orders, Thysanura and Collembola. The addition of 13 new sequences from Archaeognatha and Diplura doubles the number of apterygote insect orders with known Hox gene sequences. The Archaeognatha Hox gene sequences possibly present the best available roots for Hox genes in Hexapoda, allowing a reference point for estimations on the speed of sequence evolution of Hox genes in insects (Casillas et al., '06). In general, the new data provide a starting point for phylogenetic and developmental studies investigating the apterygote-apterygote transition.

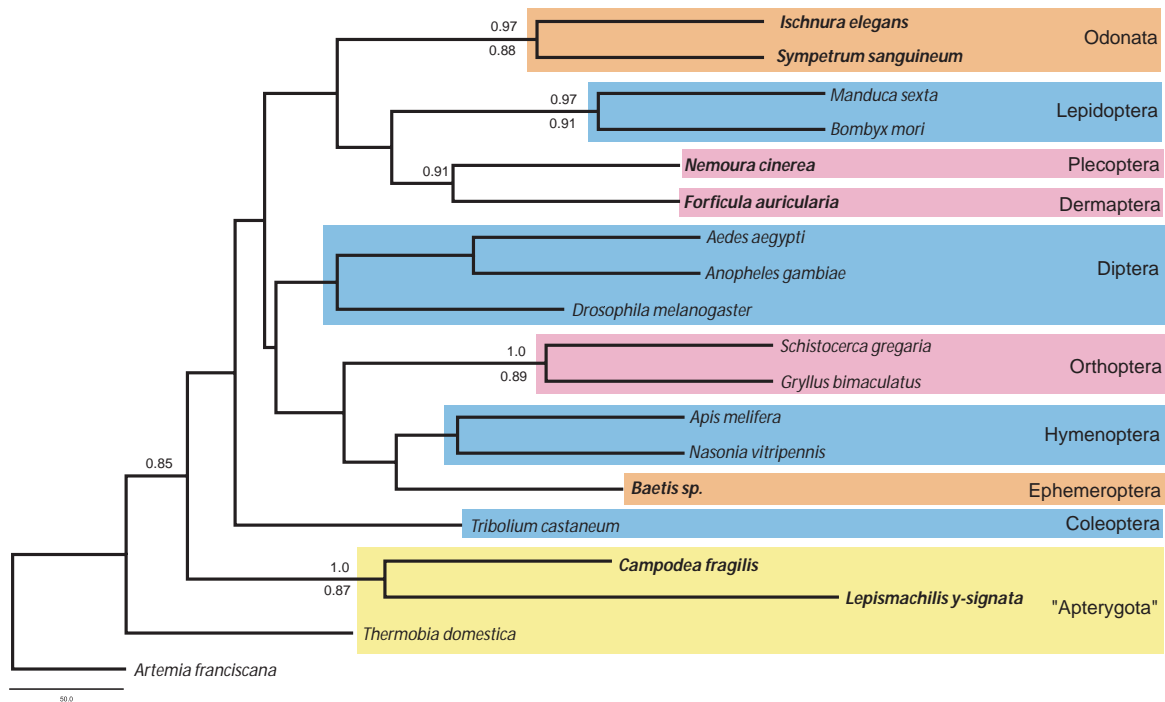
With respect to the Pterygota a large number of Hox gene sequences have previously been known from the more derived Diptera and Hymenoptera and a smaller number also from Orthoptera, Hemiptera, Coleoptera (complete cluster!) and Lepidoptera (Yasukochi et al., '04; Rogers et al., '97; Brown et al., '02). With 24 new sequences from Ephemeroptera, Odonata, Plecoptera, and Dermaptera we here add new sequences particularly from some phylogenetically crucial basal insect orders. These data are particularly important to address the origin of pterygote insects, i.e. the invention and radiation of an insect bauplan armed with wings. Most recent molecular phylogenetic analyses suggest a basal position for Odonata within the Pterygota (Simon et al., '09), making odonates particularly important for unraveling the evolutionary and developmental origin of insect wings. We could isolate all 8 of the Hox genes for odonates as well as the homeotic gene Hox3 (*bicoid*). Only the other homeotic, but non-Hox gene, *ftz*, escaped our survey. Although we increased the number of pterygote insect orders with known Hox gene sequences from 6 to 10, there is still some 19 insect orders left for which no information on Hox gene sequences is available (see Table 3).

The main goal of our study was to add as many new Hox cluster gene sequences from crucial insect orders to the database as possible. The primer pairs used in this study proved very successful for all 10 genes, but they did not amplify all homeobox fragments from all insect orders subjected to this study. Filling these gaps will obviously require a different approach and possibly different primer sets. With respect to preparing the grounds for comparative studies on the evolution of the winged insect bauplan the genes *Scr*, *Antp* and *Ubx* are of immediate importance (e.g. Ronshaugen et al., '02; Deutsch, '05; Angelini and Kaufman, '05). We have isolated fragments from all three of these genes from Archaeognatha, Ephemeroptera, Odonata, and Plecoptera. If Odonata should represent the most basal pterygote insects (see above) the new sequences from odonates will become crucial for comparative studies on the evolution of Pterygota. For this reason

we have verified the correct assignment of the new gene fragments to their *Scr*, *Antp*, and *Ubx* gene families also by RACE-PCR, amplifying full length homeobox sequences for developmental studies. These data will be reported elsewhere (Hadrys et al., in prep.).

Hox genes have been of outstanding importance for understanding the genetic mechanisms for the development of metazoan bauplans. From the very beginning of embryogenesis they control axes formation and the resulting body structuring in Bilateria (for controversial discussion on diploblastic animals see Kamm et al. ('06), Schierwater et al. ('08) and Ryan et al. ('07) and refs. therein). This developmental aspect has offered tremendous insights into details of bilaterian development in some well known model systems. The real value of these data is, however, coming from a comparative point of view, and the EvoDevo research is urgently seeking to obtain comparative data from other, non-model, animal systems, since most of the established model systems are phylogenetically quite derived and often ill-suited to address the evolutionary origin of many key bauplan transitions. One example is the invention of wings in insects, which created the unchallenged radiation success of pterygote insects. Yet this most influential evolutionary invention remains unresolved with respect to the inventor (insect order) and the genetic mechanisms (gene regulation). From some higher pterygote insects we know that for example *Scr* and *Ubx* play key roles for the formation of wings (e.g. Rogers et al., '97; Weatherbee et al., '99; Deutsch, '05; Tomoyasu et al., '05; Chesebro et al., '09) but in the absence of comparative data from more basal pterygote insect orders and in the absence of knowing the basal insect order no conclusions on the origin of the insect wing can be drawn. The new sequences from several crucial insect orders provide a first step towards obtaining the missing data. While gene expression and gene function studies will provide the relevant developmental information, phylogenetic studies have to resolve the phylogenetic relationships at the pterygote-apterygote transition.

To which degree Hox genes can also directly contribute to phylogenetic analyses has been controversially discussed (Cook et al., '01). The genomic organization of Hox genes has supported several important clades at higher taxonomic levels (e.g. Prohaska and Stadler, '06; Kamm and Schierwater, '06; Mulley et al., '06). At the sequence level of the homeobox or homeodomain one may also find phylogenetic signals at lower taxonomic levels (e.g. Casillas et al., '06; Pernice et al., '06). The main limitation here arises from the shortness of the sequence while the main strength arises from the completely unproblematic alignment (cf. Schierwater et al., '02).



**Figure 2:** The topology of Maximum Likelihood and Bayesian tree, respectively, of is identical. The matrix contains 1620 characters. Branch lengths are from maximum likelihood analysis. Bootstrap values (ML analysis) are shown below and Bayesian posterior probabilities (values below 70% are not shown) above the branches. Sequences from this study are in bold.

Our analysis of partial homeobox sequences of four Hox genes indicates that if homeobox sequences are pooled from several Hox genes, the concatenated data set may add to concatenated analyses for phylogenetic studies in insects. In the tree shown in Figure 2 the concatenated (and widely partial) homeobox sequences resolve the different insect orders. It is way beyond the scope of this paper to resolve phylogenetic relationships in insects. However, the analyses presented here clearly suggests that Hox gene homeoboxes may provide useful sequence markers that could be added to phylogenetic analyses of insect relationships.

### Acknowledgments

We thank Wolfgang Jakob for assistance and helpful comments on the manuscript. Special thanks go to Max. This work was supported by a German Research Foundation (DFG) special priority program "Deep Metazoan Phylogeny" SP1174 grant given to H.H. (DFG HA 1947/5).

**Supporting grant information:** German Research Foundation (DFG) special priority program "Deep Metazoan Phylogeny" SP1174 (DFG HA 1947/5).

### Literature Cited

- Angelini DR, Kaufman TC. 2005. Comparative developmental genetics and the evolution of arthropod body plans. *Annu Rev Genet* 39:95-119.
- Beeman RW, Stuart JJ, Brown SJ, Denell RE. 1993. Structure and function of the homeotic gene complex (HOM-C) in the beetle, *Tribolium castaneum*. *Bioessays* 15(7):439-444.
- Brown SJ, Fellers JP, Shippy TD, Richardson EA, Maxwell M, Stuart JJ, Denell RE. 2002. Sequence of the *Tribolium castaneum* homeotic complex: the region corresponding to the *Drosophila melanogaster* antennapedia complex. *Genetics* 160(3):1067-1074.
- Casillas S, Negre B, Barbadilla A, Ruiz A. 2006. Fast sequence evolution of Hox and Hox-derived genes in the genus *Drosophila*. *BMC Evol Biol* 6:106.
- Chesebro J, Hrycaj S, Mahfooz N, Popadic A. 2009. Diverging functions of Scr between embryonic and post-embryonic development in a hemimetabolous insect, *Oncopeltus fasciatus*. *Dev Biol* 329(1):142-151.
- Cook CE, Smith ML, Telford MJ, Bastianello A, Akam M. 2001. Hox genes and the phylogeny of the arthropods. *Curr Biol* 11(10):759-763.
- Deutsch J. 2005. Hox and wings. *Bioessays* 27(7):673-675.
- Duncan I. 1987. The bithorax complex. *Annu Rev Genet* 21:285-319.
- Felsenstein J. 2008. PHYLIP (phylogeny inference package), version 3.6.8. Seattle: University of Washington.
- Garcia-Fernandez J. 2005. Hox, ParaHox, ProtoHox: facts and guesses. *Heredity* 94(2):145-152.
- Hadrys H, Balick M, Schierwater B. 1992. Applications of random amplified polymorphic DNA (RAPD) in molecular ecology. *Mol Ecol* 1(1):55-63.
- Hadrys H, Schierwater B, Dellaporta SL, DeSalle R, Buss LW. 1993. Determination of paternity in dragonflies by Random Amplified Polymorphic DNA fingerprinting. *Mol Ecol* 2(2):79-87.
- Higgins DG, Sharp P.M. 1989. Fast and sensitive multiple sequence alignments on a microcomputer. *Comput Appl Biosc* 5:151 - 153.
- Jakob W, Schierwater B. 2007. Changing hydrozoan bauplans by silencing Hox-like genes. *PLoS ONE* 2(1):e694.
- Kamm K, Schierwater B. 2006. Ancient complexity of the non-Hox ANTP gene complement in the anthozoan *Nematostella vectensis*: implications for the evolution of the ANTP superclass. *J Exp Zool B Mol Dev Evol* 306(6):589-596.
- Kamm K, Schierwater B, Jakob W, Dellaporta SL, Miller DJ. 2006. Axial patterning and diversification in the cnidaria predate the Hox system. *Curr Biol* 16(9):920-926.
- Marden JH, O'Donnell BC, Thomas MA, Bye JY. 2000. Surface-skimming stoneflies and mayflies: the taxonomic and mechanical diversity of two-dimensional aerodynamic locomotion. *Physiol Biochem Zool* 73(6):751-764.
- McGinnis W, Krumlauf R. 1992. Homeobox genes and axial patterning. *Cell* 68(2):283-302.
- Mulley JF, Chiu CH, Holland PW. 2006. Breakup of a homeobox cluster after genome duplication in teleosts. *Proc Natl Acad Sci U S A* 103(27):10369-10372.
- Ogishima S, Tanaka H. 2007. Missing link in the evolution of Hox clusters. *Gene* 387(1-2):21-30.

- Panfilio KA, Akam M. 2007. A comparison of Hox3 and Zen protein coding sequences in taxa that span the Hox3/zen divergence. *Dev Genes Evol*.
- Peel AD, Telford MJ, Akam M. 2006. The evolution of hexapod engrailed-family genes: evidence for conservation and concerted evolution. *Proc Biol Sci* 273(1595):1733-1742.
- Pernice M, Deutsch JS, Andouche A, Boucher-Rodoni R, Bonnaud L. 2006. Unexpected variation of Hox genes' homeodomains in cephalopods. *Mol Phylogenet Evol* 40(3):872-879.
- Prohaska SJ, Stadler PF. 2006. Evolution of the vertebrate ParaHox clusters. *J Exp Zool B Mol Dev Evol* 306(5):481-487.
- Rogers BT, Peterson MD, Kaufman TC. 1997. Evolution of the insect body plan as revealed by the Sex combs reduced expression pattern. *Development* 124(1):149-157.
- Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572-1574.
- Ronshaugen M, McGinnis N, McGinnis W. 2002. Hox protein mutation and macroevolution of the insect body plan. *Nature* 415(6874):914-917.
- Ryan JF, Mazza ME, Pang K, Matus DQ, Baxevanis AD, Martindale MQ, Finnerty JR. 2007. Pre-Bilaterian Origins of the Hox Cluster and the Hox Code: Evidence from the Sea Anemone, *Nematostella vectensis*. *PLoS ONE* 2:e153.
- Sanchez-Herrero E, Vernos, I., Marco, R., Morata, G. 1985. Genetic organization of *Drosophila bithorax* complex. *Nature* 313:108-113.
- Schierwater B, Dellaporta S, DeSalle R. 2002. Is the evolution of Cnox-2 Hox/ParaHox genes "multicolored" and "polygenealogical?" *Mol Phylogenet Evol* 24(3):374-378.
- Schierwater B, Kamm K, Srivastava M, Rokhsar D, Rosengarten RD, Dellaporta SL. 2008. The early ANTP gene repertoire: insights from the placozoan genome. *PLoS ONE* 3(8):e2457.
- Scott MP. 1993. A rational nomenclature for vertebrate homeobox (HOX) genes. *Nucleic Acids Res* 21(8):1687-1688.
- Simon S, Strauss S, von Haeseler A, Hadrys H. 2009. A multi-gene approach using expressed sequence tags to resolve basal pterygote divergence. submitted.
- Swalla BJ. 2006. Building divergent body plans with similar genetic pathways. *Heredity* 97(3):235-243.
- Tomoyasu Y, Wheeler SR, Denell RE. 2005. Ultrabithorax is required for membranous wing identity in the beetle *Tribolium castaneum*. *Nature* 433(7026):643-647.
- Weatherbee SD, Nijhout HF, Grunert LW, Halder G, Galant R, Selegue J, Carroll S. 1999. Ultrabithorax function in butterfly wings and the evolution of insect wing patterns. *Curr Biol* 9(3):109-115.
- Yasukochi Y, Ashakumary LA, Wu C, Yoshido A, Nohata J, Mita K, Sahara K. 2004. Organization of the Hox gene cluster of the silkworm, *Bombyx mori*: a split of the Hox cluster in a non-*Drosophila* insect. *Dev Genes Evol* 214(12):606-614.



**Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships**

**Björn M von Reumont\*<sup>1</sup>, Karen Meusemann<sup>1</sup>, Nikolaus U Szucsich<sup>2</sup>, Emiliano Dell'Ampio<sup>2</sup>, Vivek Gowri-Shankar, Daniela Bartel<sup>2</sup>, Sabrina Simon<sup>3</sup>, Harald O Letsch<sup>1</sup>, Roman R Stocsits<sup>1</sup>, Yun-xia Luan<sup>4</sup>, Johann Wolfgang Wägele<sup>1</sup>, Günther Pass<sup>2</sup>, Heike Hadrys<sup>3,5</sup> and Bernhard Misof<sup>6</sup>**

<sup>1</sup>*Molecular Lab, Zoologisches Forschungsmuseum A. Koenig, Bonn, Germany*

<sup>2</sup>*Department of Evolutionary Biology, University Vienna, Vienna, Austria*

<sup>3</sup>*ITZ, Ecology & Evolution, Stiftung Tierärztliche Hochschule Hannover, Hannover, Germany*

<sup>4</sup>*Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, PR China*

<sup>5</sup>*Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA*

<sup>6</sup>*UHH Biozentrum Grindel und Zoologisches Museum, University of Hamburg, Hamburg, Germany*

\* Corresponding author

This is the author's version of a work originally published by BioMed Central Ltd in *BMC Evolutionary Biology* 2009, **9**:119; available under doi:10.1186/1471-2148-9-119

**Abstract**

**Background:** Whenever different data sets arrive at conflicting phylogenetic hypotheses, only testable causal explanations of sources of errors in at least one of the data sets allow us to critically choose among the conflicting hypotheses of relationships. The large (28S) and small (18S) subunit rRNAs are among the most popular markers for studies of deep phylogenies. However, some nodes supported by this data are suspected of being artifacts caused by peculiarities of the evolution of these molecules. Arthropod phylogeny is an especially controversial subject dotted with conflicting hypotheses which are dependent on data set and method of reconstruction. We assume that phylogenetic analyses based on these genes can be improved further i) by enlarging the taxon sample and ii) employing more realistic models of sequence evolution incorporating non-stationary substitution processes and iii) considering covariation and pairing of sites in rRNA-genes.

**Results:** We analyzed a large set of arthropod sequences, applied new tools for quality control of data prior to tree reconstruction, and increased the biological realism of substitution models. Although the split-decomposition network indicated a high noise content in the data set, our measures were able to both improve the analyses and give causal explanations for some incongruities mentioned from analyses of rRNA sequences. However, misleading effects did not completely disappear.

**Conclusion:** Analyses of data sets that result in ambiguous phylogenetic hypotheses demand for methods, which do not only filter stochastic noise, but likewise allow to differentiate phylogenetic signal from systematic biases. Such methods can only rely on our findings regarding the evolution of the analyzed data. Analyses on independent data sets then are crucial to test the plausibility of the results. Our approach can easily be extended to genomic data, as well, whereby layers of quality assessment are set up applicable to phylogenetic reconstructions in general.

## **Background**

Most recent studies that focused on the reconstruction of ancient splits in animals, have relied on 18S and/or 28S rRNA sequences, e.g. [1]. These data sets strongly contributed to our knowledge of relationships, however, several nodes remain that are suspected of being artifacts caused by peculiar evolutionary rates which may be lineage specific. Particular unorthodox nodes were discussed as long branch artifacts, others were held to be clusters caused by non-stationary evolutionary processes as indicated by differences in nucleotide composition among the terminals. The reconstruction of ancient splits seems to be especially dependent on taxon sampling and character choice, since in single lineages the signal-to-noise ratio is consistently marginal in allowing a reasonable resolution. Thus, quality assessment of data via e.g. secondary structure guided alignments, discarding of randomly similar aligned positions or heterogeneity of the data set prior to analysis is a crucial step to obtain reliable results. Arthropod phylogeny is especially suitable as a case study, since their ancient and variable phylogenetic history, which may have included intermittent phases of fast radiation, impedes phylogenetic reconstruction.

## ***Major arthropod relationships***

While currently there is wide agreement about the monophyly of Arthropoda, relationships among the four major subgroups (Chelicerata, Myriapoda, Crustacea, Hexapoda) remain contested, even the monophyly of each of the subgroups has come under question. The best supported relationship among these subgroups seems to be the clade comprising all crustaceans and hexapods. This clade, named Pancrustacea [2], or Tetraconata [3], is supported by most molecular analyses, e.g. [1, 4-14]. Likewise, the clade has increasingly found support from morphological data [3, 15-18], especially when malacostracans are directly compared with insects. Most of these studies reveal that crustaceans are paraphyletic with respect to a monophyletic Hexapoda. However, most analyses of mitochondrial genes question hexapod monophyly [19-22]. Additionally, various crustacean subgroups are discussed as potential hexapod sister groups. Fanenbruck et al. [15] favored a derivation of Hexapoda from a common ancestor with Malacostraca + Remipedia based on neuroanatomical data. In recent molecular studies, either Branchiopoda [12] or Copepoda [1, 11, 23] emerged as the sister group of Hexapoda. The Pancrustacea hypothesis implies that Atelocerata (Myriapoda + Hexapoda) is not monophyletic. In most of the above mentioned molecular studies, the Myriapoda appear at the base of the clade Mandibulata or as the sistergroup to Chelicerata. The combination of

Chelicerata + Myriapoda [1, 7, 13, 14, 24] was coined Paradoxopoda [11] or Myriochelata [10]. It seems that this grouping can be partly explained by signal erosion [25], and likewise is dependent on outgroup choice [26]. In addition, the most recent morphological data is consistent with the monophyly of Mandibulata [27], but not of Myriochelata. Almost no morphological data corroborate Myriochelata except for a reported correspondence in neurogenesis [28]; this however alternatively may reflect the plesiomorphic state within Arthropoda [29, 30]. Within Hexapoda, relationships among insect orders are far from being resolved [31-35]. Open questions concern the earliest splits within Hexapoda, e.g. the monophyly or paraphyly of Entognatha (Protura + Diplura + Collembola) [9, 19, 22, 32, 34, 36-45].

### ***Goals and methodological background***

The aim of the present study is to optimize the phylogenetic signal contained in 18S and 28S rRNA sequences for the reconstruction of relationships among the major arthropod lineages. A total of 148 arthropod taxa representing all major arthropod clades including onychophorans and tardigrades (the latter as outgroup taxa) were sampled to minimize long-branch artifacts [25]. A new alignment procedure that takes secondary structure into account is meant to corroborate the underlying hypotheses of positional homology as accurately as possible. A new tool for quality control optimizes the signal-to-noise ratio for the final analyses. In the final step, we try to improve the analyses by fitting biologically realistic mixed DNA/RNA substitution models to the rRNA data. Time-heterogeneous runs were performed to allow for lineage specific variation of the model of evolution. The use of secondary structure information both corroborates hypotheses of positional homology in the course of sequence alignment, as well as helps to avoid misleading effects of character dependence due to covariation among sites. It was demonstrated that ignoring correlated variance may mislead tree reconstructions biased by an over-emphasis of changes in paired sites [34, 46, 47]. Evolutionary constraints on rRNA molecules are well known, for example constraints resulting from secondary structure interactions. The accuracy of rRNA comparative structure models [48-50] has been confirmed by crystallographic analyses [51, 52]. Based on this background knowledge, rRNA sequences are an ideal test case to study the effect of biologically realistic substitution models on tree reconstructions. Recent studies of genome scale data revealed that a careful choice of biologically realistic substitution models and model fitting are of particular importance in phylogenetic reconstructions [53-55]. The extent, however, to which biological processes can/should be

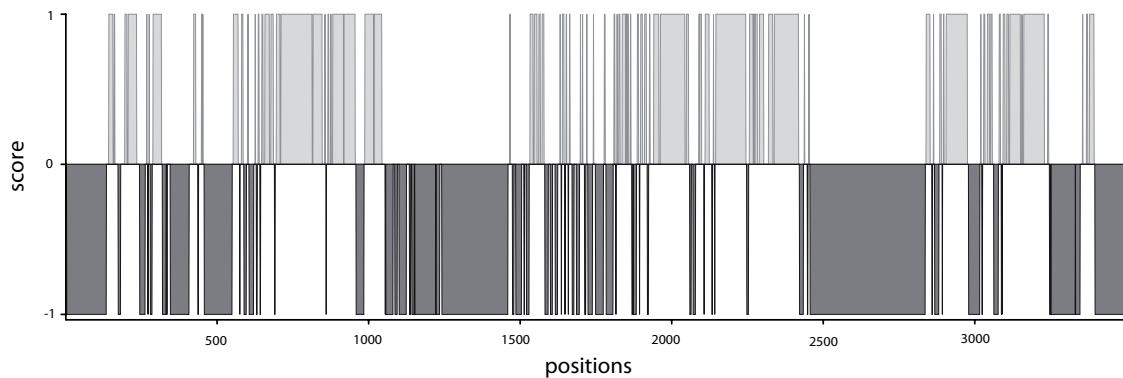
modeled in detail is still unclear. The analyses of rRNA sequences can still deliver new insights in this direction, since the relatively comprehensive background knowledge allows to better separate different aspects of the substitution processes. In order to model covariation in rRNA sequences, we estimated secondary structure interactions by applying a new approach implemented in the software RNAsalsa [56] (download available from <http://rnasalsa.zfmk.de/>), which helps to accommodate inadequate modeling (e.g. missing covariotide effects) of rRNA substitution processes in deep phylogenetic inference [34, 57]. Essentially, this approach combines prior knowledge of conserved site interactions modeled in a canonical eukaryote secondary structure consensus model with the estimation of alternative and/or additional site interactions supported by the specific data. Inferred site covariation patterns were used then to guide the application of mixed substitution models in subsequent phylogenetic analyses. Finally, we accounted for inhomogeneous base composition across taxa, a frequently observed phenomenon indicating non-stationary substitution processes [58-60]. Non-stationary processes, if present, clearly violate assumptions of stationarity regularly assumed in phylogenetic analyses [60-62]. Thus, we modeled non-stationary processes combined with the application of mixed DNA/ RNA substitution models in a Bayesian approach using the *PHASE-2.0* software package [63] to provide a better fit to our data than standard substitution models [60, 64]. In *PHASE-2.0* a nonhomogeneous substitution model is implemented [...] "by introducing a reversible jump Markov chain Monte Carlo method for efficient Bayesian inference of the model order along with other phylogenetic parameters of interest" [60]. Application of a new hierarchical prior leads to more reasonable results when only a small number of lineages share a particular substitution process. Additionally *PHASE-2.0* includes specialized substitution models for RNA genes with conserved secondary structure [60].

## Results

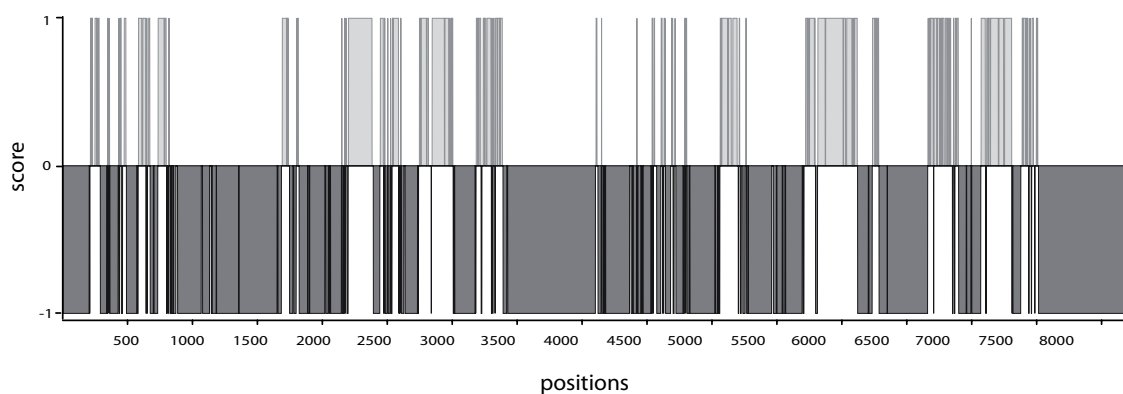
We contributed 103 new and nearly complete 18S or 28S rRNA sequences and analyzed sequences for 148 taxa (Additional file 1), of which 145 are Arthropoda sensu stricto, two onychophorans and *Milnesium* sp. (Tardigrada). The alignment of the 18S rRNA sequences comprised 3503 positions, and the 28S rRNA alignment 8184. The final secondary consensus structures included 794 paired positions in the 18S and 1326 paired positions in the 28S. The consensus structures contained all paired sites that in 60% or more sequences were detected after folding (default  $s3 = 0.6$  in RNAsalsa). ALISCORE [65]

scored 1873 positions as randomly similar (negative scoring values in the consensus profile) to the 18S and 5712 positions of the 28S alignment (Figure 1).

Aliscore profile of 18S rRNA



Aliscore profile of 28S rRNA



**Figure 1**

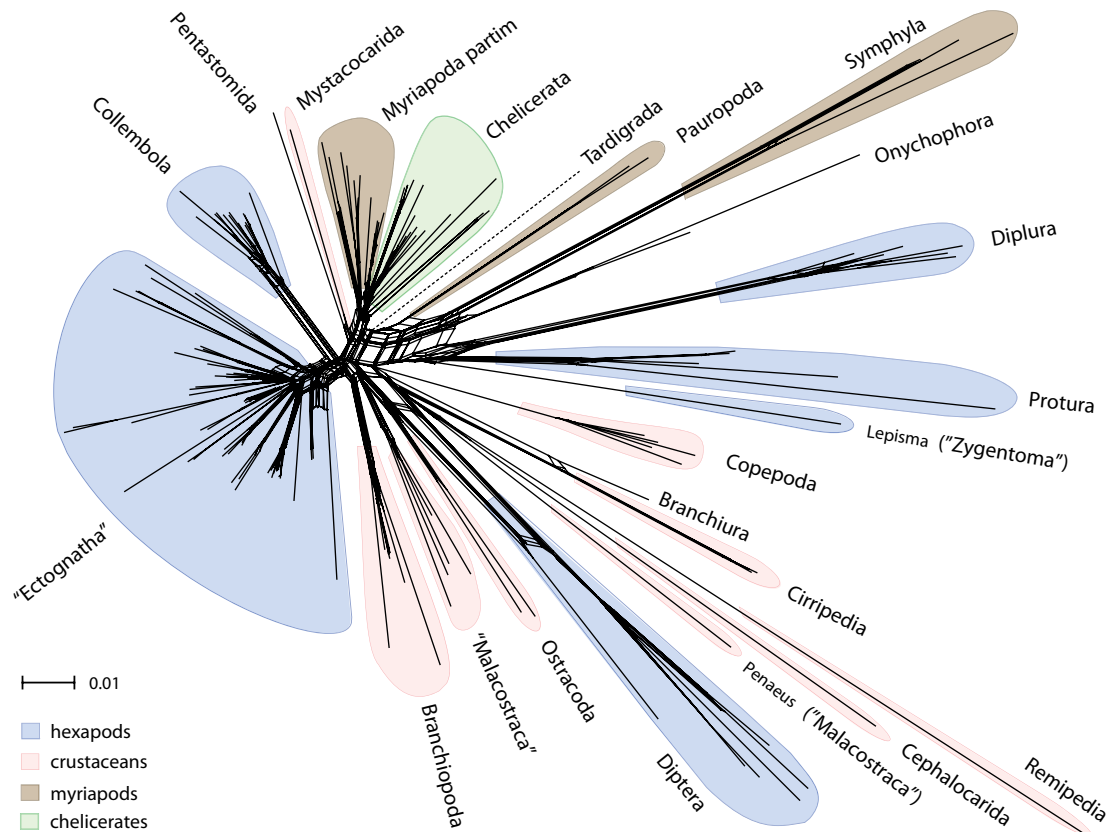
**ALISCORE consensus profiles of rRNA alignments. 1A** ALISCORE consensus profile of the 18S rRNA alignment generated from single profiles of aligned positions after applying the sliding window approach based on MC resampling. Randomly similar sections (1873 positions) show negative score values or positive values non-random similarity (y-axis). Sequence length and positions are given on the x-axis. **1B** ALISCORE consensus profile of the 28S rRNA alignment generated from single profiles of aligned positions after applying the sliding window approach based on MC resampling. Randomly similar sections (5712 positions) show negative score values or positive values for non-random similarity (y-axis). Sequence length and positions are given on the x-axis.

### *Alignment filtering and concatenation of data*

After the exclusion of randomly similar sections identified by ALISCORE, 1630 (originally 3503) of the 18S rRNA and 2472 (originally 8184) positions of the 28S rRNA remained. Filtered alignments were concatenated and used for analyses in *PHASE-2.0*. The concatenated alignment comprised 4102 positions.

***Split supporting patterns***

The neighbor-net graph, which results from a split decomposition based on uncorrected p-distances (Figure 2) and LogDet correction plus invariant sites model (see Additional file 2) pictured a dense network, which hardly resembles a tree-like topology. This indicates the presence of some problems typical in studies of deep phylogeny: a) Some taxa like Diptera (which do not cluster with ectognathous insects), Diplura, Protura and Collembola each appear in a different part of the network with Diplura and Protura separated from other hexapods, *Lepisma saccharina* (clearly separated from the second zygentoman Ctenolepisma that is nested within Ectognatha), Symphyla, Pauropoda, as well as Remipedia and Cephalocarida have very long branches. Consequently the taxa may be misplaced due to signal erosion or occurrence of homoplasies, and their placement in trees must be discussed critically [25]. The usage of the LogDet distance adjusts the length of some branches but does not decrease the amount of conflicts in deep divergence splits. b) The inner part of the network shows little treeness, which indicates a high degree of conflicting signal. A remarkable observation seen in both phylogenetic networks is that some taxa have long stem-lineages, which means that the species share distinct nucleotide patterns not present in other taxa. Such well separated groups are Copepoda, Branchiopoda, Cirripedia, Symphyla, Collembola, Diplura, Protura and Diptera, while e.g. Myriapoda partim, Chelicerata and the Ectognatha (bristletails, silverfish/firebrats and pterygote insects) excluding Diptera share weaker patterns.

**Figure 2**

**NeighborNet graph of the concatenated 18S and 28S rRNA alignment.** NeighborNet graph of the concatenated 18S and 28S rRNA alignment. NeighborNet graph based on uncorrected pdistances constructed in SplitsTree4 using the concatenated 18S and 28S rRNA alignment after exclusion of randomly similar sections evaluated with ALISCORE. Hexapods are colored blue, crustaceans red, myriapods brown and chelicerates green. Quotation marks indicate that monophyly is not supported in the given neighborNet graph.

### ***Compositional heterogeneity of base frequency***

We excluded in *PAUP* 4.0b10 [66] parsimony uninformative positions explicitly for the base compositional heterogeneity test. Randomly similar alignment blocks identified by ALISCORE were excluded for both, the base compositional heterogeneity test and phylogenetic reconstructions. 901 characters of the 18S rRNA and 1152 characters of the 28S rRNA were separately checked for inhomogeneous base frequencies. Results led to a rejection of the null hypothesis ( $H_0$ ), which assumes homogeneous base composition among taxa (18S:  $\chi^2 = 1168.94$ ,  $df = 441$ ,  $P = 0.00$ ; 28S:  $\chi^2 = 1279.98$ ,  $df = 441$ ,  $P = 0.00$ ). Thus, base frequencies significantly differed across taxa in both 18S and 28S data sets.

A data partition into stems and loops revealed 477 unpaired positions and 424 paired positions in the 18S, and 515 unpaired and 637 paired positions in the 28S. Separate



analyses of all four partitions confirmed heterogeneity of base frequencies across taxa in all sets ( $P = 0.00$  in all four partitions).

We repeated the homogeneity test for partitions as used in tree reconstruction, if base pairs were disrupted by the identification of the corresponding partner as randomly similar (ALISCORE), remaining formerly paired positions were treated as unpaired. Hence, 1848 characters of the concatenated alignment (18S: 706; 28S: 1142) were treated as paired in all analyses. Again the test revealed heterogeneity in unpaired characters of both the 18S and 28S ( $P = 0.00$  for both genes; 18S: 506 characters; 28S: 567 characters). Examination at paired positions also rejected the null hypothesis  $H_0$  (18S, 395 characters included:  $P < 0.0003$ , 28S, 585 characters included:  $P = 0.00$ ). Since non-stationary processes in all tests were strongly indicated, we chose to apply time-heterogeneous models to account for lineage-specific substitution patterns. To fix the number of "free base frequency sub-models" in time-heterogeneous analyses, we identified the minimal exclusive set of sequence groups. Based on  $\chi^2$ -tests the dataset could be divided into three groups for both rRNA genes. In both genes Diptera are characterized by a high A/T content and Diplura by a low A/T content. Exclusion of only one of the groups was not sufficient to retain a homogeneous data set (18S: excluding Diptera:  $\chi^2 = 972.91$ ,  $df = 423$ ,  $P = 0.00$ , excluding Diplura:  $\chi^2 = 532.13$ ,  $df = 423$ ,  $P < 0.0003$ ; 28S: excluding Diptera:  $\chi^2 = 986.72$ ,  $df = 423$ ,  $P = 0.00$ , excluding Diplura:  $\chi^2 = 813.8$ ,  $df = 423$ ,  $P = 0.00$ ). Simultaneous exclusion of both groups led to acceptance of  $H_0$  for 18S sequences ( $\chi^2 = 342.22$ ,  $df = 405$ ,  $P = 0.99$ ). For the 28S, after exclusion of both groups,  $H_0$  was still rejected ( $\chi^2 = 524.98$ ,  $df = 405$ ,  $P < 0.0001$ ). After sorting taxa according to base frequencies in ascending order, additional exclusion of *Peripatus* sp. and *Sinentomon erythranum* resulted in a homogeneous base composition for the 28S gene ( $H_0$ :  $\chi^2 = 434.99$ ,  $df = 399$ ,  $P = 0.1$ ), likewise indicating that three sub-models are sufficient to cover the taxon set. We repeated the homogeneity-test for stem and loop regions of each gene separately. The exclusion of Diplura was sufficient to obtain homogeneity in the loop regions for both genes (18S: 474 characters,  $P = 0.9757$ ; 28S: 541 characters,  $P = 0.0684$ ). For stem regions in the 18S it likewise was sufficient to exclude either Diptera (378 characters,  $P = 0.6635$ ) or Diplura (385 characters,  $P = 0.99$ ). These partitions would make two sub-models sufficient to cover the data set. However, in the stem regions of the 28S homogeneity was received only after the exclusion of both Diptera and Diplura (547 characters,  $P = 0.99$ ). Since PHASE-2.0 does not allow to vary

the number of chosen sub-models among partitions, we applied and fitted three sub-models to each data partition.

### ***Phylogenetic reconstructions***

Three combinations of mixed DNA/RNA models (REV +  $\Gamma$  & RNA16I +  $\Gamma$ , TN93 +  $\Gamma$  & RNA16J +  $\Gamma$  and HKY85 +  $\Gamma$  & RNA16K +  $\Gamma$ ) were compared to select the best model set. Overall model  $\ln$  likelihoods converged for all tested mixed models after a burn-in of 250,499 generations in an initial pre-run of 500,000 generations. However, most parameters did not converge for the combined REV +  $\Gamma$  & RNA16I +  $\Gamma$  models, consequently, this set up was excluded from further analyses. For each of the remaining two sets a chain was initiated for 3 million generations, with a burn-in set to 299,999 generations. The applied Bayes Factor Test [[67, 68], BFT], favored the TN93 +  $\Gamma$  & RNA16J +  $\Gamma$  model combination ( $2\ln B_{10} = 425.39$ , harmonic mean  $\ln L_0$  (TN93 +  $\Gamma$  & RNA16J +  $\Gamma$ ) = 79791.08; harmonic mean  $\ln L_1$  (HKY85 +  $\Gamma$  & RNA16K +  $\Gamma$ ) = 80003.78). For each approach (Additional file 3) all chains which passed a threshold value in a BFT were assembled to a metachain. Each resulting extended majority rule consensus tree was rooted with *Milnesium*. Node support values for clades were deduced from 56,000 sampled trees for the time-heterogeneous set (Figure 3) and from 18,000 sampled trees for the time-homogeneous set (Figure 4), detailed support values are shown in Additional file 3. Harmonic means of the  $\ln$  likelihoods of included time-heterogeneous chains were compared against all  $\ln$  likelihoods of included time-homogeneous chains (burnin discarded) in a final BFT: the time-heterogeneous model was strongly favored ( $2\ln B_{10} = 1362.13$ ).

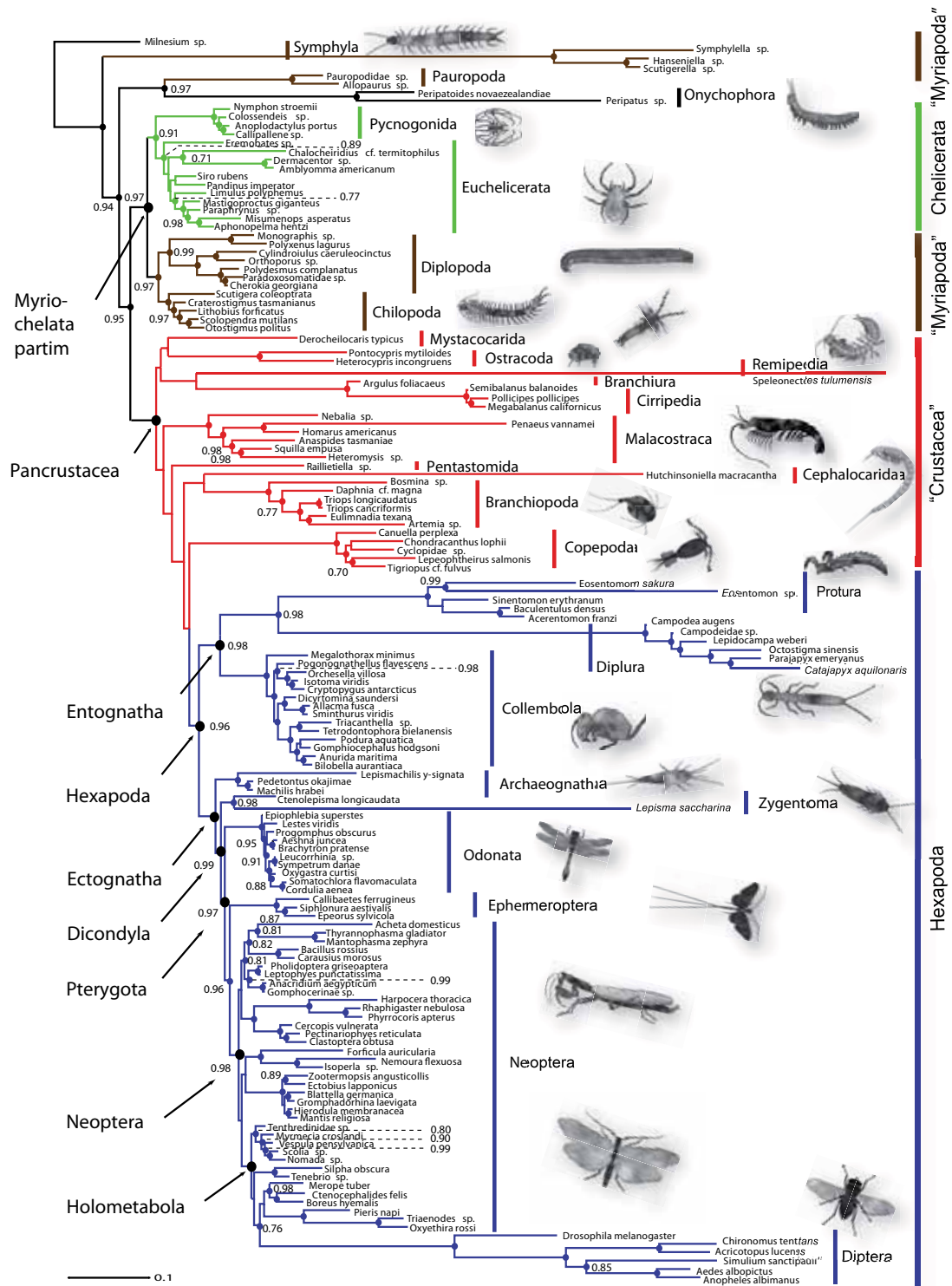
### ***Resulting topologies***

Representatives of Symphyla and Pauropoda, already identified in the neighbor-net graph as taxa with conspicuously long branches (Figure 2), assumed unorthodox positions in both trees which are clearly incongruent with morphological evidence and results obtained from other genes. Symphyla formed the sister group of all remaining arthropod clades, and Pauropoda clustered with Onychophora. Consequently, myriapods always appeared polyphyletic in both analyses. We consider these results as highly unlikely, since they contradict all independent evidence from morphology, development, and partly from other genes. In the following, we focus on major clades and point out differences between time-heterogeneous tree (Figure 3) and time-homogeneous tree (Figure 4) without considering

the position of Symphyla and Pauropoda. Possible causes for the misplacement of these groups, however, will be treated in the discussion. Both analyses supported a monophyletic Chelicerata (pP 0.91 in the time-heterogeneous tree and maximal support in the time-homogeneous tree) with Pycnogonida (sea spiders) as sister group to remaining chelicerates. Pycnogonida received maximal support in both analyses. Euchelicerata received highest support in the time-homogeneous approach while this clade in the time-heterogeneous approach received a support of only pP 0.89. *Limulus polyphemus* (horseshoe crab) clustered within arachnids, but some internal relationships within Euchelicerata received only low support. Chilopoda always formed the sister group of a monophyletic Diplopoda in both analyses with high support. Within the latter the most ancient split lied between Penicillata and Helminthomorpha. This myriapod assemblage – Myriapoda partim – formed the sister group of Chelicerata, thus giving support to the Myriochelata hypothesis, respectively Myriochelata partim, when the long-branch clades Symphyla and Pauropoda are disregarded.

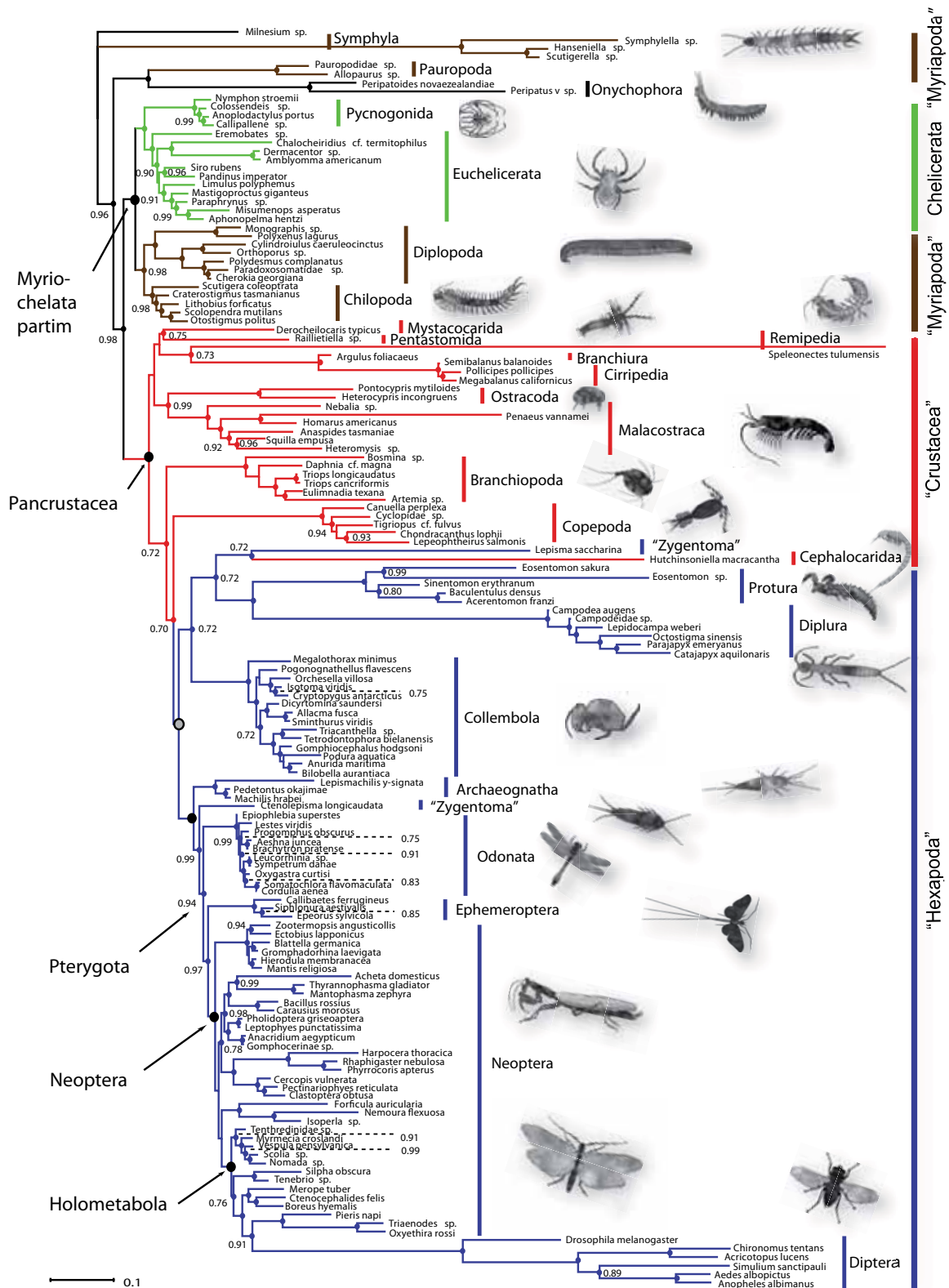
Pancrustacea showed always maximal support. The monophyly of Malacostraca and Branchiopoda received highest support in both approaches while their position varied. Branchiopoda was the sister group of the clade consisting of Copepoda + Hexapoda in the homogeneous tree (Figure 4), however the cephalocarid *Hutchinsoniella* nested within hexapods. Among hexapods, monophyly was unambiguously supported for Protura, Diplura, Collembola, Archaeognatha, Odonata, Ephemeroptera, Phasmatodea, Mantophasmatodea, Mantodea, Plecoptera, Hemiptera, Coleoptera, Hymenoptera, Lepidoptera and Diptera. Diplura clustered with Protura, and gave support to a monophyletic Nonoculata. Pterygota occurred in both topologies, well supported in the non-stationary tree (pP 0.97) and with moderate support (pP 0.94) in stationary tree. Within the winged insects, both analyses resolved Odonata as the sister group to a well supported monophyletic clade Ephemeroptera + Neoptera (heterogeneous: pP 0.96; homogeneous: pP 0.97), known as the "Chiasmomyaria" clade [32, 34, 35, 69]. Blattodea were always paraphyletic with respect to the isopteran representative. This assemblage formed a sister group relationship with Mantodea, thus giving support to a monophyletic Blattopteroidea or Dictyoptera while the position of Dictyoptera among hemimetabolan insects differed. Dermaptera always clustered with Plecoptera. Hemiptera (Heteroptera + Homoptera) in both approaches formed a clade with the remaining orthopterans + ((*Acheta* + Mantophasmatodea)Phasmatodea) with low statistical support. Caused by *Acheta*

orthopteran insects appeared always polyphyletic. Within the mono-phyletic Holometabola (pP 1.0), Hymenoptera formed the sister group of the remaining taxa.



**Figure 3**

**Time-heterogeneous consensus tree.** Consensus tree from 56,000 sampled trees of the time-heterogeneous substitution process inferred by *PHASE-2.0*, graphically processed with Adobe Illustrator CS2. Support values below 0.70 are not shown (nodes without dots), nodes with a maximum posterior probability (pP) of 1.0 are represented by dots only. Quotation marks indicate that monophyly is not supported in the given tree.

**Figure 4**

**Time-homogeneous consensus tree.** Consensus tree from 18,000 sampled trees of the time-homogeneous substitution process inferred by *PHASE-2.0*, graphically processed with Adobe Illustrator CS2. Support values below 0.70 are not shown (nodes without dots), nodes with a maximum posterior probability (pP) of 1.0 are represented by dots only. The grey dot indicates the clade containing all hexapod taxa including Hutchinsoniella (Crustacea) + Lepisma (Zygentoma); its node value is pP 0.58. Quotation marks indicate that monophyly is not supported in the given tree.

While the time-heterogeneous and time-homogeneous trees corresponded in overall topologies, they differed in a number of remarkable details.

1) Hexapoda, Entognatha, Ectognatha and Dicondylia were only reconstructed in the time-heterogeneous approach. 2) The cephalocarid *Hutchinsoniella* clustered among crustaceans as sister group to the Branchiopoda only in the heterogeneous approach, this clade formed the sister group to (Copepoda + Hexapoda) although with low support. 3) The time-homogeneous runs revealed highly supported (Malacostraca + Ostracoda) as the sister group to a clade ((Mystacocarida + Pentastomida) + (Branchiura + Cirripedia)). In contrast, in the time-heterogeneous analysis more terminal positioned Malacostraca are the sister group of a clade (Pentastomida((Cephalocarida + Branchiopoda) + (Copepoda + Hexapoda))). The altered position of Pentastomida was only low supported in this tree. 4) In the homogeneous tree *Hutchinsoniella* emerged as sister taxon to *Lepisma* with low support (pP 0.72), and this cluster was positioned within the remaining hexapods (Figure 4). Hexapoda were monophyletic only in the time-heterogeneous approach, well supported (pP 0.96, Figure 3), with Copepoda as sister group, latter with low support (pP 0.69). 5) In the time-homogeneous tree (Figure 4), Copepoda emerged as sister group, again with a low support value (pP 0.70) of ((*Lepisma* + *Hutchinsoniella*) + "Hexapoda"). 6) Entognatha (pP 0.98), and Ectognatha (pP 1.0) and Dicondylia (pP 0.99) were monophyletic only in the time-heterogeneous tree. 7) The time-heterogeneous tree showed the expected paraphyly of primarily wing-less insects with Archaeognatha as sister group to Zygentoma + Pterygota. 8) Within pterygote insects (Dermaptera + Plecoptera) emerged as sister group of Dictyoptera in the non-stationary tree, contrary as sister group of Holometabola in the stationary tree, both scenarios with negligible support.

## Discussion

Among arthropods 18S and 28S rRNA genes have the densest coverage of known sequences. Apart of some exceptions most studies on phylogenetic relationships at least partly rely on rRNA data. Often, however, only one of the genes was used, sometimes even just fragments of a gene [23, 32, 34, 40, 42, 44, 70-72], while only few studies used nearly complete 18S and 28S rRNA sequences [1, 11, 73]. Despite this wide usage, the reliability of reconstructions based on rRNA markers is still debated (for contradicting views see [34, 74, 75]. A major cause of concern is the pronounced site heterogeneity of evolutionary

rates, the non-stationarity of base composition among taxa and rate variation in time. All three phenomena quickly lead to the erosion of phylogenetic signal [76]. On the one hand, our understanding of the molecular structure of other markers and about taxon-dependent processes of molecular evolution remains poor. On the other hand, our vast background knowledge regarding rRNA molecules offers a unique opportunity to study the effects of selection and application of substitution models in greater detail.

### ***Quality check and character choice in alignments***

Phylogenetic signal in sequence data can get noisy due to (i) multiple substitution processes (saturation) and (ii) erroneous homology hypotheses caused by ambiguous sequence alignment. Both effects correspond in that they result in random similarity of alignment regions. Such noisy sections potentially bias tree reconstructions in several ways which have been appreciated for years but only recently been applied, that allow to account for these problems [25, 54, 77, 78]. Exclusion of these ambiguously aligned or saturated regions can help to reduce noise, see e.g. [65]. If this topic is addressed at all, the majority of studies include a manual alignment check for untrustworthy regions [1, 4, 22, 32, 34, 39, 44, 71-73]. Only some recent publications addressing arthropod relationships have used automated tools, e.g. [14, 79, 80].

To identify alignment sections of random similarity prior to tree reconstructions, we used ALISCORE, which, compared to the commonly used Gblocks [81], is not dependent on the specification of an arbitrary threshold [65]. To improve the signal-to-noise ratio we restricted our character choice to alignment sections which contained nucleotide patterns that differ from randomized patterns.

### ***Phylogenetic reconstruction methods***

Arthropod phylogenies have been inferred with reconstruction methods like Maximum Parsimony, Maximum Likelihood and Bayesian approaches. We tried to implement knowledge about the evolution of rRNA in two ways: (i) the use of mixed DNA/RNA models is meant to account for known instances of character dependence due to compensatory mutations in stem regions, (ii) the application of time-heterogeneous models accounts for non-stationary processes that occurred in arthropod lineages. The consensus secondary structure of our dataset, generated with RNAsalsa, can be understood as a model parameter that defines site interactions and thus character dependence due to compensatory

mutations [34, 82, 83]. Neglect of character dependence surely results in unrealistic support values. In single low supported nodes, where the signal-to-noise ratio is at the edge of resolution, such a neglect theoretically can even turn the balance between two competing hypotheses. Additionally a consensus secondary structure is necessary to apply a mixed model approach, since it determines whether the evolution of a given site is modeled by the DNA-model, or as part of a base-pair by the RNA-model. Within the mixed model approach, we opted for DNA-corresponding 16-state RNA models [63]. It can certainly be argued that the choice of 16-state models is problematic because it is difficult to fit these models to real data due to their parameter richness and heavy computational costs. However, even the best choice of a consensus secondary structure can only capture the predominantly conserved structural features among the sequences. This implies that the applied RNA models must be able to cope with mismatches in base-pairing. Less complex RNA models like those of the 6 and 7-state families either ignore mismatches completely or pool these mismatches into a single character state which produces artificial synapomorphies. Additionally, according to Schöninger and v. Haeseler [84], it is more likely that co-variation is a multiple step process which allows for the intermediate existence of instable (non Watson-Crick) pairs. These intermediate states are only described in 16-state RNA models.

Concerning rRNA-genes of arthropods, shifts in base composition are mentioned for Diptera, Diplura, Protura and Symphyla [1, 23, 34, 44, 73, 85]. Since base compositional heterogeneity within a dataset can mislead phylogenetic reconstruction [61, 86, 87] and [60], some of these studies discussed observed but not incorporated non-stationary processes as possible explanations for misplacements of some taxa [11, 23, 24, 44, 73]. The selective exclusion of these taxa to test for misleading effects on the remaining topology, however, is not appropriate to test whether non-stationarity really fits as the causal explanation of the placement incongruent with other analyses. LogDet methods have been applied to compensate for variations of base frequencies [1, 11, 44], which leads to some independence of non-stationarity, but among site rate variation (ASRV) cannot be handled efficiently. After detecting compositional base frequency heterogeneity in our data, we chose a non-stationary approach implemented in *PHASE-2.0*. Because no previous study of arthropod phylogeny has used a time-heterogeneous approach including mixed DNA/RNA models, we compared this approach with a "classical" time-homogeneous setup. Our results prove that the time-heterogeneous approach produces



improved likelihood values with improved branch lengths estimates and more realistic, though not perfect (see below), topology estimates. Since modeling of general time-heterogeneous processes is in its infancy and since its behavioural effect on real data is relatively unknown [60, 61], we favored a set up accounting for the three different "submodels" corresponding to three base frequency categories in our dataset (Additional file 4). The application of the three submodels to individual branches in a tree by the MCMC process was not further constrained. This scheme allowed for a maximum of flexibility without losing the proper mix of parameters.

### ***Conflicting phylogenetic hypotheses and non-stationary processes of rRNA evolution***

The comparison of our time-homogeneous approach to our time-heterogeneous one was not only meant to show improvements in the application of more realistic models, but also to indicate which incongruities of analyses of rRNA genes may be causally explained by non-stationary processes during the evolution of these genes.

In our time-homogeneous approach, the crustacean *Hutchinsoniella* (Cephalocarida) clustered with *Lepisma* (Zygentoma, Hexapoda) within entognathans as sister group to Nonoculata (Protura + Diplura), (see Figure 4). This led to the polyphyly or paraphyly of several major groups (e.g. Hexapoda, Entognatha, Ectognatha, Dicondylia). In our time-heterogeneous analysis, Cephalocarida clustered as sister group to Branchiopoda. This result, although marginal supported, is congruent, at least, with some morphological data [88]. Most recent molecular studies have not included Cephalocarida, e.g. [1, 11]. Regier et al. [12] reconstructed a sister group relationship of Remipedia and Cephalocarida (likewise represented by *Hutchinsoniella*), but his result also received only moderate bootstrap support. The same clade was presented in Giribet et al. [9] based on morphological and molecular data.

Independent of the sister group relationship of Cephalocarida within crustaceans, the correction of the misplacement of *Hutchinsoniella*, by allowing for non-stationary processes, has a major effect on the heuristic value of our analyses. Not only is the monophyletic status of Hexapoda, Entognatha, Ectognatha, Dicondylia supported after the correction, but likewise a causal explanation is given for the misplacement in the time-homogeneous approach, which cannot be accomplished by alternatively excluding the taxon. Our time-heterogeneous analyses resulted in a sister group relationship of Diplura

and Protura, which lends support to a monophyletic Nonoculata within a monophyletic Entognatha. This result is congruent with trees published by Kjer [32], Luan et al. [44], Mallat and Giribet [1], and Dell'Ampio et al. [23]. Following Luan et al. [44] Dell'Ampio et al. [23] cautioned that Nonoculata may be an artificial cluster caused by a shared nucleotide bias and long branch attraction. Since this node is recovered with high support by our non-stationary approach, Nonoculata cannot be suspected of being an artificial group based on shared compositional biases alone. However, one must keep in mind that Protura and Diplura have longer branches than Ectognatha and Collembola (Figure 3 and 4), and long-branch effects may still be present. Thus monophyly of a clade Nonoculata still awaits support from a data set independent from rRNA sequences.

### ***Clades not affected by non-stationary processes***

#### ***Symphyla and Pauropoda***

Although we tried to break down long branches by a dense taxon sampling, some long-branch problems persisted. We cannot clearly address the reason but, due to the symptoms, assume that saturation by multiple substitution caused signal erosion (class II effect, [25]). To evaluate the impact on the topology of the very likely incorrect positions of Symphyla and Pauropoda, we repeated the time-heterogeneous analysis using a reduced dataset excluding these taxa. We limited the analysis to ten chains with 7, 000, 000 generations each (2, 000, 000 burn-in). Differences occurring in the inferred consensus topology (not shown) of the final three chains (15, 000, 000 generations) show that some nodes are still sensitive to taxon sampling, since e.g. Pycnogonida clustered with (Chilopoda + Diplopoda) after exclusion of pauropod and symphylian sequences. Also the crustacean topology changed, remaining long branch taxa *Hutchinsoniella* and *Speleonectes* clustered together in the reduced dataset, forming a clade with (Branchiura + Cirripedia).

#### ***Mandibulata versus Myriochelata***

Analyses of rRNA sequences up till now were held to favor Myriochelata (Myriapoda + Chelicerata) over Mandibulata [1, 4, 11]. Our analyses provide no final conclusion with respect to this conflict, since the position of Pauropoda and Symphyla is unusual, it results in polyphyletic myriapods. The exact reconstruction of the position of myriapods within the Euarthropoda thus demands e.g. the application of new markers and suitable phylogenetic strategies.

*Phylogenetic position of Malacostraca and Pentastomida*

The position of Malacostraca differs among molecular studies. Often, Malacostraca emerge as nested within the remaining crustacean groups, e.g. [5, 89]. Complete mitochondrial genomes place Malacostraca close to insects [90, 91]. However, studies of rRNA sequences recover this group as the sister group to all remaining crustaceans [1, 11, 92]. Since in our stationary tree monophyletic Malacostraca branched off at a more basal split within crustaceans [88, 93], forming a sister group relationship to Ostracoda and contrary they branched off at a more terminal split in the non-stationary tree we cannot draw a final conclusion about the placement of Malacostraca. Unfortunately the position of the Pentastomida remains ambiguous in our analyses, we argue that low pP values might be induced by conflicting phylogenetic signal.

*Sister group of Hexapoda*

The sister group of Hexapoda is still disputed. Most molecular studies support paraphyly of crustaceans with respect to hexapods. A sister group relationship between Branchiopoda and Hexapoda was proposed for the first time by Regier and Shultz [94], yet with low support. Shultz and Regier [5] and Regier et al. [12] corroborated this relationship, which is likewise favored by authors of rRNA-based studies [1, 11], despite their result that Cyclopidae (Copepoda) is the sister group of Hexapoda. Our denser taxon sampling further supports Copepoda as the sister group to Hexapoda, but the low support value might indicate conflicting signal. This clade up till now, however, lacks any support from morphological studies.

*Ancient splits within pterygote insects*

We find that the rRNA data cannot robustly resolve the most ancient splits within Pterygota. Nonetheless, rRNA data, when analyzed under more realistic models favour Chiasmomyaria as the most likely hypothesis. Since all three possible arrangements of Odonata, Ephemeroptera and Neoptera likewise receive morphological support, we agree with Whitfield and Kjer [35] that the ambiguity can best be explained by early 'explosive radiation' within Pterygota.

## Conclusion

We conclude that the implementation of biologically realistic model parameters, such as site interaction (mixed DNA/RNA models) and compositional heterogeneity of base frequency, is fundamental to robustly reconstruct phylogenies. The most conspicuous examples comparing our trees are a) the position of *Hutchinsoniella* (Crustacea), although a low pP value of 0.59 in the non-stationary tree prohibits conclusions about its internal crustacean relationship and b) the well supported position of *Ctenolepisma* and *Lepisma* (Zygentoma). As a consequence, the monophyly of Hexapoda, Entognatha and Ectognatha and Dicondylia received support only in the time-heterogeneous approach. Several artificial clades remain in our analyses which cannot be causally explained unambiguously. However, the examples given here clearly demonstrate that the probability to causally explain some incongruities between different data sets, as well as the correction of certain obvious misplacements, is enhanced by using more complex but realistic models. The present study aimed to incorporate background knowledge on the evolution of molecular sequences in general and ribosomal RNA-genes in special into various steps of data processing. For all steps fully automated methods were used, including an automated secondary structure guided alignment approach, a software that enables to distinguish random similarity from putative phylogenetic signal, mixed models that avoid artefacts due to co-variation among sites, and analyses that account for variation of evolutionary rates among lineages. The resolution of many relationships among arthropods, and the minimization of obvious misplacements demonstrate that the increased computational effort pays off.

## Methods

### *Taxon Sampling*

Our taxon sampling was designed to represent a taxonomically even collection of specimens across arthropod groups. In particular, we took care to include taxa which do not differ too widely from the hypothetical morphological ground-pattern of the represented group, when possible [53, 78]. In total we included 148 concatenated 18S and 28S rRNA sequences in the analysis (Additional file 1). Of these, we contributed 103 new sequences, 41 for the 18S and 62 for the 28S rRNA gene, respectively. Only sequences which span at least 1500 bp for the 18S and 3000 bp for the 28S were included. For 29 taxa we had to construct chimeric concatenated sequences of 18S and 28S rRNA sequences of different species, marked with an asterisk. Details are listed in Additional file 5, we chose

species as closely related as possible depending on its availability in GenBank. The outgroup included the concatenated 18S and 28S rRNA sequences of *Milnesium* sp. (Tardigrada).

### **Laboratory work**

Collected material was preserved in 94 – 99% ethanol or liquid nitrogen. Samples were stored at temperatures ranging from -20°C to -80°C. DNA extraction of complete specimens or tissue followed different standard protocols. We used phenol-chloroform isoamyl extraction [95], standard column DNA extraction kits DNeasy Blood & Tissue Kit (Qiagen) and NucleoSpin Tissue Kit (Machery-Nagel) following the manual. Single specimens were macerated for extraction, only specimens of *Ctenocephalides felis* were pooled. Manufacturer protocols were modified for all crustaceans, some apterygote hexapods and myriapods (overnight incubation and adding 8 µl RNase [10 mg/ml] after lysis). Extracted genomic DNA was amplified with the Illustra GenomiPhi V2 DNA Amplification Kit (GE Healthcare) for tiny, rare or hard to collect specimens.

Partly published rRNA primer sets were used, they were designed in part for specific groups (Additional file 6 and 7). The 18S of crustaceans was amplified in one PCR product and sequenced using four primer combinations. The 18S of apterygotes was amplified in three or four fragments (Additional file 8). The 28S of crustaceans and basal hexapods was amplified in nine overlapping fragments starting approximately in the middle of the rRNA 5.8S to the nearly end of the D12 of 28S rRNA (Additional file 9). The 28S of odonats was amplified in seven or eight, the 28S of ephemeropterans and neopterans in eight overlapping fragments (Additional file 10). Primers were ordered from Metabion, Biomers or Sigma-Genosys. PCR products were purified using following kits: NucleoSpin ExtractionII (Machery-Nagel), QIAquick PCR purification kit (Qiagen), peqGOLD Gel Extraction Kit (peqLab Biotechnologie GmbH), MultiScreen PCR Plate (Millipore) and ExoI (Biolabs Inc.)/SAP (Promega). Some samples were purified using a NHAc [4 mol] based ethanol precipitation. In case of multiple bands fragments with the expected size were cut from 1% – 1.5% agarose gel and purified according to manufacturer protocols.

Cycle sequencing and sequence analyses took place on different thermocyclers and sequencers. Cycle sequencing products were purified and sequenced double stranded. Several amplified and purified PCR products were sequenced by MacroGen (Inc.), Korea.

Sequencing of the 28S fragment 28V – D10b.PAUR of the Pauropodidae sp. (Myriapoda) was only successful via cloning. Fragments of the 28S rRNA of the diplopod *Monographis* sp. (Myriapoda) were processed following Mallatt et al. [11] and Luan et al. [44]. Please refer to the electronic supplement (Additional file 11) for detailed information about PCR-conditions, applied temperature profiles (Additional file 12), primer combinations, used chemicals (Additional file 13) and settings to amplify DNA fragments. Sequence electropherograms were analyzed and assembled to consensus sequences applying the software SeqMan (DNASStar Lasergene) or BioEdit 7.0 [96]. All sequences or composed fragments were blasted in NCBI using BLASTN, mega BLAST or "BLAST 2 SEQUENCES" [97] to exclude contaminations.

### ***Alignments and alignment evaluation***

Secondary structures of rRNA genes were considered (as advocated in [98-101]) to improve sequence alignment. Structural features are the targets of natural selection, thus the primary sequence may vary, as long as the functional domains are structurally retained. Alignments and their preparation for analyses were executed for each gene separately. We prealigned sequences using MUSCLE v3.6 [102]. Sequences of 24 taxa of Pterygota were additionally added applying a profile-profile alignment [103]. The 28S sequences of *Hutchinsoniella macracantha* (Cephalocarida), *Speleonectes tulumensis* (Remipedia), *Raillietiella* sp. (Pentastomida), *Eosentomon* sp. (Protura) and *Lepisma saccharina* (Zygentoma) were incomplete. Apart from *L. saccharina*, prealignments of these taxa had to be corrected manually. We used the "BLAST 2 SEQUENCES" tool to identify the correct position of sequence fragments in the multiple sequence alignment (MSA) for these incomplete sequences.

The software RNAsalsa [56] is a new approach to align structural RNA sequences based on existing knowledge about structure patterns, adapted constraint directed thermodynamic folding algorithms and comparative evidence methods. It automatically and simultaneously generates both individual secondary structure predictions within a set of homologous RNA genes and a consensus structure for the data set. Successively sequence and structure information is taken into account as part of the alignment's scoring function. Thus, functional properties of the investigated molecule are incorporated and corroborate homology hypotheses for individual sequence positions. The program employs a progressive multiple alignment method which includes dynamic programming and affine

gap penalties, a description of the exact algorithm of RNA-salsa will be presented elsewhere.

As a constraint, we used the 28S + 5.8S (U53879) and 18S (V01335) sequences and the corresponding secondary structures of *Saccharomyces cerevisiae* extracted from the European Ribosomal Database [104-106]. Structure strings were converted into dot-bracket-format using Perlscripts. Folding interactions between 28S and 5.8S [74, 107, 108] required the inclusion of the 5.8S in the constraint to avoid artificial stems. Alignment sections presumably involved in the formation of pseudoknots were locked from folding to avoid artifacts. Pseudoknots in *Saccharomyces cerevisiae* are known for the 18S (stem 1 and stem 20, V4-region: stem E23 9, E23 10, E23 11 and E23 13) while they are lacking in the 28S secondary structure. Prealignments and constraints served as input, and RNAsalsa was run with default parameters. We constructed manually chimeran 18S sequences of *Speleonectes tulumensis* (EU370431, present study and L81936) and 28S sequences of *Raillietiella* sp. (EU370448, present study and AY744894). Concerning the 18S of *Speleonectes tulumensis* we combined positions 1–1644 of L81936 and positions 1645–3436 of sequence EU370043. Regarding the 28S of *Raillietiella* we combined positions 1–3331 of AY744894 with positions 3332–7838 of sequence EU370448. Position numbers refer to aligned positions.

RNAsalsa alignments were checked with ALISCORE [65]. ALISCORE generates profiles of randomness using a sliding window approach. Sequences within this window are assumed to be unrelated if the observed score does not exceed 95% of scores of random sequences of similar window size and character composition generated by a Monte Carlo resampling process. ALISCORE generates a list of all putative randomly similar sections. No distinction is made between random similarity caused by mutational saturation and alignment ambiguity. A sliding window size ( $w = 6$ ) was used, and gaps were treated as ambiguities (- N option).

The maximum number of possible random pairwise comparisons (- r: 10,878) was analyzed. After the exclusion of putative random sections and uninformative positions using *PAUP* 4.0b10, alignments were checked for compositional base heterogeneity using the .2-test. Additionally, for each sequence the heterogeneity-test was performed for paired

and unpaired sites separately. Further heterogeneity-tests were applied to determine the minimal number of base frequency groups.

RNAseal generated consensus structure strings for 18S and 28S rRNA sequences, subsequently implemented in the MSA. Randomly similar sections identified by ALISCORE were excluded using a Perl-script. ALISCORE currently ignores base pairings. If ambiguously aligned positions within stems are discarded the corresponding positions will be handled as an unpaired character in the tree reconstruction. The cleaned 18S and 28S alignments were concatenated.

To analyze information content of raw data SplitsTree4 was used to calculate phylogenetic networks (see Huson and Bryant [109] for a review of applications). We compared the network structure based on the neighbor-net algorithm [110] and applying the LogDet transformation, e.g. [111, 112]. LogDet is a distance transformation that corrects for biases in base composition. The network graph gives a first indication of signal-like patterns and conflict present in the alignments. We used the alignment after filtering of random-like patterns with ALISCORE.

### ***Phylogenetic reconstruction***

Mixed DNA/RNA substitution models were chosen, in which sequence partitions corresponding to loop regions were governed by DNA models and partitions corresponding to stem regions by RNA models that consider co-variation. Among site rate variation [113] was implemented in both types of substitution models. Base frequency tests indicated that base composition was inhomogeneous among taxa (see results), suggesting non-stationary processes of sequence evolution. To take such processes into account the analyses were performed in *PHASE-2.0* [63] to accommodate this compositional heterogeneity to minimize bias in tree reconstruction. Base compositional heterogeneity is implemented in PHASE2.0 according to the ideas developed by Foster [87].

We limited the number of candidate models to the REV +  $\Gamma$ , TN93 +  $\Gamma$  and the HKY85 +  $\Gamma$  models for loop regions and the corresponding RNA16I +  $\Gamma$ , RNA16J +  $\Gamma$  and RNA16K +  $\Gamma$  models for stem regions. Site heterogeneity was modeled by a discrete gamma distribution [114] with six categories. The extent of invariant characters was not estimated since it was shown to correlate strongly with the estimation of the shape parameter of the



gamma distribution [113, 115-117]. The data was partitioned into four units representing loop and stem regions of 18S rRNA and loop and stem regions of 28S rRNA. DNA and RNA substitution model parameters were independently estimated for each partition. Substitution models were selected based on results of time-homogeneous setups. We tested three different combinations of substitution models, REV +  $\Gamma$  & RNA16I +  $\Gamma$ , TN93 +  $\Gamma$  & RNA16J +  $\Gamma$  and HKY85 +  $\Gamma$  & RNA16K +  $\Gamma$ . We used Dirichlet distribution for priors, proposal distribution and Dirichlet priors and proposals for a set of exchangeability parameters (Additional file 14) described in Gowri-Shankar and Rattray [60].

Appropriate visiting of the parameter space according to the posterior density function [118] was checked by plotting values of each parameter and monitoring their convergence. This was calculated for all combinations after 500,000 generations (sampling period: 150 generations). We discarded models in which values of several parameters did not converge. For models which displayed convergence of nearly all parameter values, we re-run MCMC processes with 3,000,000 generations and a sampling period of 150 generations. Prior to comparison of the harmonic means of  $\ln L$  values, 299,999 generations were discarded as burn-in. After a second check for convergence the model with the best fitness was selected applying a Bayes Factor Test (BFT) to the positive values of the harmonic means calculated from  $\ln L$  values [67, 68]. The favored model ( $2\ln B_{10} > 10$ ) was used for final phylogenetic reconstructions.

To compare results of time-homogeneous and time-heterogeneous models, 14 independent chains of 7,000,000 generations and two chains of 10 million generations for both setups were run on Linux clusters (Pentium 4, 3.0 GHz, 2 Gb RAM, and AMD Opteron Dual Core, 64 bit systems, 32 Gb RAM). For each chain the first two million generations were discarded as burn-in (sampling period of 1000). The setup for the time-homogeneous approach was identical to the pre-run except for number of generations, sampling period and burn-in. The setting for the time-heterogeneous approach differed (Additional file 4). We followed the method of Foster [87] and Gowri-Shankar and Rattray [60] in the non-homogeneous setup whereby only a limited number of composition vectors can be shared by different branches in the tree. Exchangeability parameters (average substitution rate ratio values, rate ratios and alpha shape parameter) were fixed as input values. Values for these parameters were computed from results of the preliminary time-homogeneous pre-run (3,000,000 generations). A consensus tree was inferred in PHASE *mcmcsummarize*

using the output of the pre-run. This consensus tree topology and the model file of this run served as input for a ML estimation of parameters in PHASE *optimizer*. Estimated values of exchangeability parameters from the resulting *optimizer* output file and estimated start values for base frequencies were fed into *mcmcphase* for the time-heterogeneous analysis. Values of exchangeability parameters remained fixed during the analysis. The number of allowed base frequency categories (models) along the tree was also fixed. The number of base frequency groups was set to three "submodels"), reflecting base frequency heterogeneity.

Harmonic means of  $\ln L$  values of these 16 independent chains were again compared with a BFT to identify possible local optima in which a single chain might have been trapped. We only merged sample data of chains with a  $2\ln B_{10}$ -value  $< 10$  [67] using a Perl-script to construct a "metachain" [119]. Finally we included ten time-heterogeneous chains and three time-homogeneous chains. The assembled meta-chains included 56 million generations for the non-stationary approach (Additional file 15) and 18 million generations for the time-homogeneous approach (Additional file 16), burn-ins were discarded. Consensus trees and posterior probability values were inferred using *mcmcsummarize*. Branch lengths of the time-homogeneous and time-heterogeneous consensus tree were estimated using three *mcmcphase* chains (4 million generations, sampling period 500, topology changes turned off, starting tree = consensus tree, burn-in: 1 million generations) from different initial states with a Gowri-Shankar modified PHASE version. To infer mean branch lengths we combined data with the described branch lengths and *mcmcsummarize*. These mean branch lengths were used to redraw the consensus tree (Additional file 4).

Localities of sampled specimen used for amplification are listed in Additional file 17.

### List of abbreviations

rRNA: ribosomal RNA; PCR: polymerase chain reaction; RNA: ribonucleic acid; DNA: deoxyribonucleic acid; df: degree of freedom; P: probability; pP: posterior probability; sp.: species epithet not known;  $\ln$ : natural logarithm or  $\log_e$ ; BFT: Bayes Factor Test.

### Authors' contributions

BMvR, KM and BM conceived the study, designed the setup and performed all analyses. VG complemented PHASE-2.0 and contributed to PHASE-2.0 analyses setup. RRS, HOL,

BM provided RNAsalsa and software support. JWW allocated the neighbor-net-analysis. BMvR, KM, ED, SS, HOL, DB and YL contributed sequence data and designed primers. BMvR, KM, BM, NUS and JWW wrote the paper with comments and revisions from ED, VG, RRS, DB, SS, GP, HH and YL. All authors read and approved the final manuscript.

## **Additional material**

### **Additional file 1**

Taxa list. Taxa list of sampled sequences. \* indicates concatenated 18S and 28S rRNA sequences from different species. For combinations of genes to construct concatenated sequences of chimeran taxa, see Table S1. \*\* contributed sequences in the present study (author of sequences).

### **Additional file 2**

LogDet corrected network of concatenated 18S and 28S rRNA alignment. LogDet corrected network plus invariant site models (30.79% invariant sites) using SplitsTree4 based on the concatenated 18S and 28S rRNA alignment after exclusion of randomly similar sections evaluated with ALISCORE.

### **Additional file 3**

Bayesian support values for selected clades. List of Bayesian support values (posterior probability, pP) for selected clades of the time-heterogeneous and time-homogeneous tree.

### **Additional file 4**

Detailed flow of the analysis procedure in the software package *PHASE-2.0*. Options used in *PHASE-2.0* are italicized above the arrows and are followed by input files. Black arrows represent general flows of the analysis procedure, green arrows show that results or parameter values after single steps were inserted or accessed in a further process. Red block-arrows mark the final run of the time-heterogeneous and time-homogeneous approach with 16 chains each ( $2 \times 118,000,000$  generations). First row: I.) We prepared 3 control files (control.mcmc) for *mcmcphase* using three different mixed models. This "pre-run" was used for a first model selection (500,000 generations for each setting). We excluded model (C) based on non-convergence of parameter values. II.) We repeated step one (I.) with 3,000,000 generations using similar control files (different number of generations and random seeds) of the two remaining model settings. Calculated *ln* likelihoods values of both chains were compared in a BFT resulting in the exclusion of mixed model (A). Parameter values of the remaining model (B) were implemented in the time-heterogeneous setting. III.) We started the final analysis (final run) using sixteen

chains for both the time-homogeneous and the time-heterogeneous approach. In the final time-homogeneous approach, the control files were similar to step II.) except for a different number of generations and random seeds. Second row: Additional steps were necessary prior to the computation of the final time-heterogeneous chains. We applied *mcmcsummarize* for the selected mixed model (B) to calculate a consensus tree. *Optimizer* was executed to conduct a ML estimation for each parameter value (opt.mod) based on the inferred consensus tree and optimized parameter-values (mcmc-best.mod), a data file delivered by *mcmcphase*. Estimated values were implemented in an initial.mod file. The initial.mod file and its parameter values were accessed by the control files of the final time-heterogeneous chains (only topology and base frequencies estimated). Third row: Trees were reconstructed separately for the time-homogeneous and time-heterogeneous setting. All chains of each approach were tested in a BFT against the chain with the best lnL. We only included chains with a  $2\ln B_{10}$ -value  $> 10$ . From these chains we constructed a metachain for each setting using Perl and applied *mcmcsummarize* to infer the consensus topology. To estimate branch lengths properly we ran *mcmcphase*, resulting branch lengths were implemented in the consensus trees. Finally, both trees were optimized using graphic programs (Dendroscope, Adobe Illustrator CS II).

#### **Additional file 5**

List of chimeran species for concatenated 18S and 28S rRNA sequences.

#### **Additional file 6**

Primer list 18S rRNA.

#### **Additional file 7**

Primer list 28S rRNA.

#### **Additional file 8**

Primercard of the 18S rRNA gene for hexapods, myriapods and crustaceans. Primers used for hexapods or myriapods are shown in the upper part, primers for crustaceans in the lower part. Positions of forward primers are marked with green arrows, those of reverse primers with red arrows. When different primers with identical position were used, all primer labels are given at the single arrow for the specific position. Primers and their combinations are given in Additional file 6 and 11.

#### **Additional file 9**

Primercard of the 28S rRNA gene for crustaceans, hexapods and myriapods. Positions of forward primers are tagged with green arrows, those of reverse primers with red arrows. When different primers with identical position were used, all primer labels are given at the

single arrow for the specific position. Primers and their combinations are given in Additional file 7 and 11.

**Additional file 10**

Primercard of the 28S rRNA gene for pterygots. Positions of forward primers are assigned by green arrows, those of reverse primers with red arrows. When different primers with identical position were used, all primer labels are given at the single arrow for the specific position. Primers and their combinations are given in Additional file 7 and 11.

**Additional file 11**

Supplementary Information. Supplementary information for lab work (amplification, purification and sequencing of PCR products).

**Additional file 12**

PCR temperature-profiles.

**Additional file 13**

PCR chemicals.

**Additional file 14**

Setting of exchangeability parameters used in pre-runs. Listed settings of exchangeability parameters used in pre-runs in *PHASE-2.0*.

**Additional file 15**

Included chains to infer the time-heterogeneous consensus tree. Number of chains, generations per chain, harmonic means ( $\ln L$ ) and  $2\ln B_{10}$ -values included to infer the time-heterogeneous consensus tree.

**Additional file 16**

Included chains to infer the time-homogeneous consensus tree. Number of chains, generations per chain, harmonic means ( $\ln L$ ) and  $2\ln B_{10}$ -values included to infer the time-homogeneous consensus tree.

**Additional file 17**

Localities of sampled taxa.

**Acknowledgements**

We thank Matty Berg, Anke Braband, Antonio Carapelli, Erhard Christian, Romano Dallai, Johannes Dambach, Wolfram Dunger, Erich Eder, Christian Epe, Pietro Paolo Fanciulli, Makiko Fikui, Francesco Frati, Peter Frenzel, Yan Gao, Max Hable, Bernhard Huber, Herbert Kliebhan, Stefan Koenemann, Franz Krabb, Ryuichiro Machida, Albert Melber, Wolfgang Moser, Reinhard Predel, Michael Raupach, Sven Sagasser, Kaoru

Sekiya, Marc Sztatecsny, Dieter Waloßek, Manfred Walzl and Yi-ming Yang for help in collecting specimens, for providing tissue or other DNA-samples or for laboratory help. Thanks also go to Andreas Wißkirchen, Theory Department, Physikalisches Institut, University of Bonn for using their computational power. We thank Berit Ullrich, Oliver Niehuis and Patrick Kück for providing Perl-Scripts and Thomas Stamm for suggestions on the discussion structure. Special thanks go to John Plant for linguistic help. This work was supported by the German Science Foundation (DFG) in the priority program SPP 1174 "Deep Metazoan Phylogeny" <http://www.deep-phylogeny.org>. Work by JWW, BMvR is supported by the DFG grant WA 530/34; BM, KM are funded by the DFG grant MI 649/6, HH and SS are supported by the DFG grant HA 1947/5. NUS, ED, DB and GP are funded by the Austrian Science Foundation (FWF) grant P 20497-B17.

## References

1. Mallatt J, Giribet G: **Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch.** *Mol Phylogenet Evol* 2006, **40**(3):772-794.
2. Zrzavy J, Stys P: **The basic body plan of arthropods: insights from evolutionary morphology and developmental biology.** *J Evol Biol* 1997, **10**(3):653-367.
3. Dohle W: **Are the insects terrestrial crustaceans? A discussion of some new facts and arguments and the proposal of the proper name "Tetraconata" for the monophyletic unit Crustacea + Hexapoda.** *Ann Soc Entomol Fr (New Series)* 2001, **37**(3):85-103.
4. Friedrich M, Tautz D: **Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods.** *Nature* 1995, **376**(6536):165-167.
5. Shultz JW, Regier JC: **Phylogenetic analysis of arthropods using two nuclear protein-encoding gene ssupports a crustacean + hexapod clade.** *Proc Biol Sci* 2000, **267**(1447):1011-1019.
6. Friedrich M, Tautz D: **Arthropod rDNA phylogeny revisited: A consistency analysis using Monte Carlo simulation.** *Ann Soc Entomol Fr (New Series)* 2001, **37**(1-2):21-40.
7. Hwang UW, Friedrich M, Tautz D, Park CJ, Kim W: **Mitochondrial protein phylogeny joins myriapods with chelicerates.** *Nature* 2001, **413**(6852):154-157.
8. Regier JC, Shultz JW: **Elongation factor-2: A useful gene for arthropod phylogenetics.** *Mol Phylogenet Evol* 2001, **20**(1):136-148.
9. Giribet G, Edgecombe GD, Wheeler WC: **Arthropod phylogeny based on eight molecular loci and morphology.** *Nature* 2001, **413**(6852):157-161.
10. Pisani D, Poling L, Lyons-Weiler M, Hedges SB: **The colonization of land by animals: molecular phylogeny and divergence times among arthropods.** *BMC Biol* 2004, **2**(1):1.
11. Mallatt JM, Garey JR, Shultz JW: **Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin.** *Mol Phylogenet Evol* 2004, **31**(1):178-191.

12. Regier JC, Shultz JW, Kambic RE: **Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic.** *Proc Biol Sci* 2005, **272**(1561):395-401.
13. Hassanin A: **Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution.** *Mol Phylogenet Evol* 2006, **38**(1):100-116.
14. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD *et al*: **Broad phylogenomic sampling improves resolution of the animal tree of life.** *Nature* 2008, **452**(7188):745-749.
15. Fanenbruck M, Harzsch S, Wägele JW: **The brain of the Remipedia (Crustacea) and an alternative hypothesis on their phylogenetic relationships.** *Proc Natl Acad Sci USA* 2004, **101**(11):3868-3873.
16. Harzsch S, Müller CHG, Wolf H: **From variable to constant cell numbers: cellular characteristics of the arthropod nervous system argue against a sister-group relationship of Chelicerata and "Myriapoda" but favour the Mandibulata concept.** *Dev Genes Evol* 2005, **215**(2):53-68.
17. Harzsch S: **Neurophylogeny: Architecture of the nervous system and a fresh view on arthropod phylogeny.** *Integr Comp Biol* 2006, **46**(2):162-194.
18. Ungerer P, Scholtz G: **Filling the gap between identified neuroblasts and neurons in crustaceans adds new support for Tetraconata.** *Proc Biol Sci* 2008, **275**(1633):369-376.
19. Nardi F, Spinsanti G, Boore JL, Carapelli A, Dallai R, Frati F: **Hexapod origins: monophyletic or paraphyletic?** *Science* 2003, **299**(5614):1887-1889.
20. Cameron SL, Miller KB, D'Haese CA, Whiting MF, Barker SC: **Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea \textit{sensu lato} (Arthropoda).** *Cladistics* 2004, **20**(6):534-557.
21. Cook CE, Yue Q, Akam M: **Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic.** *Proc Biol Sci* 2005, **272**(1569):1295-1304.
22. Carapelli A, Liò P, Nardi F, van der Wath E, Frati F: **Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea.** *BMC Evol Biol* 2007, Suppl 2:S8.
23. Dell'Ampio E, Szucsich NU, Carapelli A, Frati F, Steiner G, Steinacher A, Pass Gu: **Testing for misleading effects in the phylogenetic reconstruction of ancient lineages of hexapods: influence of character dependence and character choice in analyses of 28S rRNA sequences.** *Zool Scr* 2009, **38**(2):155-170.
24. Hassanin A, L'eger N, Deutsch J: **Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of Metazoa, and consequences for phylogenetic inferences.** *Syst Biol* 2005, **54**(2):277-298.
25. Wägele JW, Mayer C: **Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects.** *BMC Evol Biol* 2007, **7**(1):147.
26. Rota-Stabelli O, Telford ML: **A multi criterion approach for the selection of optimal outgroups in phylogeny: Recovering some support for Mandibulata over Myriochelata using mitogenomics.** *Mol Phylogenet Evol* 2008, **48**(1):103-111.
27. Bäcker H, Fanenbruck M, Wägele JW: **A forgotten homology supporting the monophyly of Tracheata: The subcoxa of insects and myriapods re-visited.** *Zool Anz* 2008, **247**(3):185-207.

28. Kadner D, Stollewerk A: **Neurogenesis in the chilopod *Lithobius forficatus* suggests more similarities to chelicerates than to insects.** *Dev Genes Evol* 2004, **214**(8):367-379.
29. Stollewerk A, Simpson P: **Evolution of early development of the nervous system: a comparison between arthropods.** *Bioessays* 2005, **27**(9):874-883.
30. Stollewerk A, Chipman AD: **Neurogenesis in myriapods and chelicerates and its importance for understanding arthropod relationships.** *Integr Comp Biol* 2006, **46**(2):195-206.
31. Ogden TH, Whiting MF: **The problem with "the Paleoptera Problem": sense and sensitivity.** *Cladistics* 2003, **19**(5):432-442.
32. Kjer KM: **Aligned 18S and insect phylogeny.** *Syst Biol* 2004, **53**(3):506-514.
33. Kukalova-Peck J, Lawrence JF: **Relationships among coleopteran suborders and major endoneopteran lineages: Evidence from hind wing characters.** *Eur J Entomol* 2004, **101**(1):95-144.
34. Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A: **Towards an 18S phylogeny of hexapods: Accounting for group-specific character covariance in optimized mixed nucleotide/doublet models.** *Zoology (Jena)* 2007, **110**(5):409-429.
35. Whitfield JB, Kjer KM: **Ancient rapid radiations of insects: Challenges for phylogenetic analysis.** *Annu Rev Entomol* 2008, **53**(1):449-472.
36. Koch M: **Monophyly and phylogenetic position of the Diplura (Hexapoda).** *Pedobiologia (Jena)* 1997, **41**(9):9-12.
37. Kristensen NP: **The groundplan and basal diversification of the hexapods.** In: *Arthropod Relationships*. London: Chapman and Hall: 281-293.
38. Carapelli A, Frati F, Nardi F, Dallai R, Simon C: **Molecular phylogeny of the apterygotan insects based on nuclear and mitochondrial genes.** *Pedobiologia (Jena)* 2000, **44**(3-4):361-373.
39. Carapelli A, Nardi F, Dallai R, Boore JL, Li'o P, Frati F: **Relationships between hexapods and crustaceans based on four mitochondrial genes.** In: *Crustacean and Arthropod Relationships*. vol. 16: CRC Press; 2005: 295-306.
40. D'Haese CA: **Were the first springtails semi-aquatic? A phylogenetic approach by means of 28S rDNA and optimization alignment.** *Proc Biol Sci* 2002, **269**(1496):1143-1151.
41. Luan Y-x, Zhang Y, Qiaoyun Y, Pang J, Xie R, Yin W: **Ribosomal DNA gene and phylogenetic relationships of Diplura and lower hexapods.** *Sci China, C, Life Sci* 2003, **46**(1):67-76.
42. Giribet G, Edgecombe GD, Carpenter JM, D'Haese CA, Wheeler WC: **Is Ellipura monophyletic? A combined analysis of basal hexapod relationships with emphasis on the origin of insects.** *Org Divers Evol* 2004, **4**(4):319-340.
43. Regier JC, Shultz JW, Kambic RE: **Phylogeny of basal hexapod lineages and estimates of divergence times.** *Ann Entomol Soc Am* 2004, **97**(9):411-419.
44. Luan Y-x, Mallatt JM, Xie R-d, Yang Y-m, Yin W-y: **The phylogenetic positions of three basal-hexapod groups (Protura, Diplura, and Collembola) based on on ribosomal RNA gene sequences.** *Mol Biol Evol* 2005, **22**(7):1579-1592.
45. Szucsich NU, Pass Gu: **Incongruent phylogenetic hypotheses and character conflicts in morphology: The root and early branches of the hexapodan tree.** *Mitt Dtsch Ges Allg Angew Entomol* 2008, **16**:415-429.
46. Jow H, Hudelot C, Rattray M, Higgs PG: **Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution.** *Mol Biol Evol* 2002, **19**(9):1591-1601.



47. Galtier N: **Sampling properties of the bootstrap support in molecular phylogeny: Influence of nonindependence among sites.** *Syst Biol* 2004, **53**(1):38-46.
48. Fox GE, Woese CR: **The architecture of 5S rRNA and its relation to function.** *J Mol Evol* 1975, **6**(1):61-76.
49. Wuyts J, De Rijk P, Van de Peer Y, Pison G, Rousseeuw P, De Wachter R: **Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA.** *Nucleic Acids Res* 2000, **28**(23):4698-4708.
50. Gutell JC, Robin R, Lee, Cannone JJ: **The accuracy of ribosomal RNA comparative structure models.** *Curr Opin Struct Biol* 2002, **12**(3):301-310.
51. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA: **The complete atomic structure of the large ribosomal subunit at 2.4 Å Resolution.** *Science* 2000, **289**(5481):905-920.
52. Noller HF: **RNA structure: reading the ribosome.** *Science* 2005, **309**(5740):1508-1514.
53. Lartillot N, Philippe He: **Improvement of molecular phylogenetic inference and the phylogeny of Bilateria.** *Philos Trans R Soc Lond, B, Biol Sci* 2008, **363**(1496):1463-1472.
54. Rodriguez-Ezpeleta N, Brinkmann H, Roure Be, Lartillot N, Lang BF, Philippe He: **Detecting and overcoming systematic errors in genome-scale phylogenies.** *Syst Biol* 2007, **56**(3):389-399.
55. Philippe He, Germot Ae, Moreira D: **The new phylogeny of eukaryotes.** *Curr Opin Genet Dev* 2000, **10**(6):596-601.
56. Stocsits RR, Letsch H, Hertel J, Misof B, Stadler PF: **RNA-salsa.** In.; 2008.
57. Brown JM, Lemmon AR: **The importance of data partitioning and the utility of bayes factors in bayesian phylogenetics.** *Syst Biol* 2007, **56**(4):643-655.
58. Galtier N, Gouy M: **Inferring phylogenies from DNA sequences of unequal base compositions.** *Proc Natl Acad Sci USA* 1995, **92**(24):11317-11321.
59. Tarrio R, Rodríguez-Trelles F, Ayala FJ: **Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae.** *Mol Biol Evol* 2001, **18**(8):1464-1473.
60. Gowri-Shankar V, Rattray M: **A reversible jump method for bayesian phylogenetic inference with a nonhomogeneous substitution model.** *Mol Biol Evol* 2007, **24**(6):1286-1299.
61. Blanquart S, Lartillot N: **A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution.** *Mol Biol Evol* 2006, **23**(11):2058-2071.
62. Gowri-Shankar V, Rattray M: **On the correlation between composition and site-specific evolutionary rate: Implications for phylogenetic inference.** *Mol Biol Evol* 2006, **23**(2):352-364.
63. Gowri-Shankar V, Jow H: **PHASE: a software package for Phylogenetics And Sequence Evolution.** In.; 2006.
64. Telford MJ, Wise MJ, Gowri-Shankar V: **Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: Examples from the Bilateria.** *Mol Biol Evol* 2005, **22**(4):1129-1136.
65. Misof B, Misof K: **A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion.** *Syst Biol* 2009, **58**(1):sy006.
66. Swofford DL: **PAUP\* Phylogenetic Analysis Using Parsimony (\*and Other Methods).** In., 4 edn. Sunderland, MA: Sinauer Associates; 2002.

67. Kaas RE, Raftery AE: **Bayes Factors**. *Journal of the American Statistical Association* 1995, **90**(430):773-795.
68. Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JeL: **Bayesian phylogenetic analysis of combined data**. *Syst Biol* 2004, **53**(21):47-67.
69. Boudreaux BH: **Arthropod phylogeny: with special reference to insects**: John Wiley & Sons Inc.; 1979.
70. Edgecombe GD, Giribet G: **Myriapod phylogeny and the relationships of Chilopoda**. In: *Biodiversidad, Taxonomía y Biogeografía de Artrópodos de México: Hacia una Síntesis de su Conocimiento*. vol. III: Aula-Verlag; 2002: 143-168.
71. Kjer KM, Carle FL, Litman J, Ware J: **A Molecular Phylogeny of Hexapoda**. *Arthropod Systematics & Phylogeny* 2006, **64**(1):35-44.
72. Yamaguchi S, Endo K: **Molecular phylogeny of Ostracoda (Crustacea) inferred from 18S ribosomal DNA sequences: implication for its origin and diversification**. *Mar Biol* 2003, **143**(1):23-38.
73. Gai Y-H, Song D-X, Sun H-Y, Zhou K-Y: **Myriapod monophyly and relationships among myriapod classes based on nearly complete 28S and 18S rDNA sequences**. *Zool Sci* 2006, **23**(12):1101-1108.
74. Gillespie JJ, Johnston JS, Cannone JJ, Gutell RR: **Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) rRNA genes of *Apis mellifera* (Insecta: Hymenoptera): structure, organization, and retrotransposable elements**. *Insect Mol Biol* 2005, **15**(5):657-686.
75. Jordal B, Gillespie JJ, Cognato AI: **Secondary structure alignment and direct optimization of 28S rDNA sequences provide limited phylogenetic resolution in bark and ambrosia beetles (Curculionidae: Scolytinae)**. *Zool Scr* 2008, **37**(1):43-56.
76. Simon C, Buckley TR, Frati F, Stewart JB, Beckenbach AT: **Incorporating Molecular Evolution into Phylogenetic Analysis, and a New Compilation of Conserved Polymerase Chain Reaction Primers for Animal Mitochondrial DNA**. *Annu Rev Ecol Evol Syst* 2006, **37**:545-579.
77. Susko E, Spencer M, Roger AJ: **Biases in phylogenetic estimation can be caused by random sequence segments**. *J Mol Evol* 2005, **61**(3):351-359.
78. Philippe He, Delsuc Fee, Brinkmann H, Lartillot N: **Phylogenomics**. *Annual Review of Ecology, Evolution, and Systematics* 2005, **36**(1):541-562.
79. Roeding F, Hagner-Holler S, Ruhberg H, Ebersberger I, Arndt vH, Kube M, Reinhardt R, Burmester T: **EST sequencing of Onychophora and phylogenomic analysis of Metazoa**. *Mol Phylogenet Evol* 2007, **45**(3):942-951.
80. Podsiadlowski L, Kohlhagen H, Koch M: **The complete mitochondrial genome of *Scutigera coleoptrata* (Myriapoda: Symphyla) and the phylogenetic position of Symphyla**. *Mol Phylogenet Evol* 2007, **45**(1):251-260.
81. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis**. *Mol Biol Evol* 2000, **17**(4):540-552.
82. Hancock JM, Tautz D, Dover GA: **Evolution of the secondary structures and compensatory mutations of the ribosomal RNAs of *Drosophila melanogaster***. *Mol Biol Evol* 1988, **5**(4):393-414.
83. Stephan W: **The rate of compensatory evolution**. *Genetics* 1996, **144**(1):419-426.
84. Schöninger M, von Haeseler A: **A stochastic model for the evolution of autocorrelated DNA sequences**. *Mol Phylogenet Evol* 1994, **3**(3):240-247.
85. Friedrich M, Tautz D: **An episodic change of rDNA nucleotide substitution rate has occurred during the emergence of the insect order Diptera**. *Mol Biol Evol* 1997, **14**(6):644-653.

86. Jermini LS, Ho SYW, Ababneh F, Robinson J, Larkum AWD: **The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated.** *Syst Biol* 2004, **53**(4):638-643.
87. Foster PG: **Modeling compositional heterogeneity.** *Syst Biol* 2004, **53**(3):485-495.
88. Walossek D: **On the Cambrian diversity of Crustacea.** In: *Crustaceans and the Biodiversity Crisis, Proceedings of the Fourth International Crustacean Congress, Amsterdam, The Netherlands, July 20-24, 1998.* vol. 1: Brill Academic Publishers, Leiden; 1998: 3-27.
89. Edgecombe GD, Wilson GDF, Colgan DJ, Gray MR, Cassis G: **Arthropod Cladistics: Combined analysis of histone H3 and U2 snRNA sequences and morphology.** *Cladistics* 2000, **16**(2):155-203.
90. Lim JT, Hwang UW: **The complete mitochondrial genome of *Pollicipes mitella* (Crustacea, Maxillopoda, Cirripedia): non-monophyly of Maxillopoda and Crustacea.** *Mol Cells* 2006, **22**(3):314-322.
91. Wilson K, Cahill V, Ballment E, Benzie J: **The complete sequence of the mitochondrial genome of the crustacean *Penaeus monodon*: Are malacostracan crustaceans more closely related to insects than to branchiopods?** *Mol Biol Evol* 2000, **17**(6):863-874.
92. Glenner H, Thomsen PF, Hebsgaard MB, Sørensen MV, Willerslev E: **Evolution: The origin of insects.** *Science* 2006, **314**(5807):1883-1884.
93. Zhang X-g, Siveter DJ, Waloszek D, Maas A: **An epipodite-bearing crown-group crustacean from the Lower Cambrian.** *Nature* 2007, **449**(7162):595-598.
94. Regier JC, Shultz JW: **Molecular phylogeny of the major arthropod groups indicates polyphyly of crustaceans and a new hypothesis for the origin of hexapods.** *Mol Biol Evol* 1997, **14**(9):902-913.
95. Schierwater B, Hadrys H: **Environmental factors and metagenesis in the hydroid *Eleutheria dichotoma*.** *Invertebr Reprod Dev* 1998, **34**(2-3):139-148.
96. Hall TA: **BioEdit: a user-friendly biological alignment sequence EDITOR and analysis program for Windows95/98/NT.** *Nucleic Acids Symp Ser* 1999, **41**(2-3):95-98.
97. Tatusova TA, Madden TL: **BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences.** *FEMS Microbiol Lett* 1999, **174**(2):247-250.
98. Kjer KM: **Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: An example of alignment and data presentation from the frogs.** *Mol Phylogenet Evol* 1995, **4**(3):314-330.
99. Hickson RE, Simon C, Perrey SW: **The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence.** *Mol Biol Evol* 2000, **17**(4):530-539.
100. Buckley TR, Simon C, Chambers GK: **Exploring among-site rate variation models in a maximum likelihood framework using empirical data: Effects of model assumptions on estimates of topology, branch lengths, and bootstrap support.** *Syst Biol* 2001, **50**(1):67-86.
101. Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A: **A hexapod nuclear SSU rRNA secondary-structure model and catalog of taxon-specific structural variation.** *J Exp Zool B Mol Dev Evol* 2006, **306B**(1):70-88.
102. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**(5):1792-1797.
103. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**(1):113.

104. Van de Peer Y, De Rijk P, Wuyts J, Winkelmans T, De Wachter R: **The European Small Subunit Ribosomal RNA database.** *Nucleic Acids Res* 2000, **28**(1):175-176.
105. Wuyts J, Perri\`ere G, Van de Peer Y: **The European ribosomal RNA database.** *Nucleic Acids Res* 2004, **32**(Suppl 1, Database issue):D101-103.
106. Wuyts J, Van de Peer Y, Winkelmans T, De Wachter R: **The European database on small subunit ribosomal RNA.** *Nucleic Acids Res* 2002, **30**(1):183-185.
107. Michot B, Bachellerie J-P, Raynal F: **Structure of mouse rRNA precursors. Complete sequence and potential folding of the spacer regions between 18S and 28S rRNA.** *Nucleic Acids Res* 1983, **11**(10):3375-3391.
108. Gillespie JJ, Munro JB, Heraty JM, Yoder MJ, Owen AK, Carmichael AE: **A secondary structural model of the 28S rRNA expansion segments D2 and D3 for chalcidoid wasps (Hymenoptera: Chalcidoidea).** *Mol Biol Evol* 2005, **22**(7):1593-1608.
109. Huson DH, Bryant D: **Application of phylogenetic networks in evolutionary studies.** *Mol Biol Evol* 2006, **23**(2):254-267.
110. Bryant D, Moulton V: **Neighbor-Net: An agglomerative method for the construction of phylogenetic networks.** *Mol Biol Evol* 2004, **21**(2):255-265.
111. Penny D, Lockhart PJ, Steel MA, Hendy MD: **The role of models in reconstructing evolutionary trees.** In: *Models in phylogeny reconstruction.* Oxford University Press; 1994: 211-230.
112. Steel M, Huson D, Lockhart PJ: **Invariable sites models and their use in phylogeny reconstruction.** *Syst Biol* 2000, **49**(2):225-232.
113. Yang Z: **Among-site rate variation and its impact on phylogenetic analyses.** *Trends Ecol Evol (Amst)* 1996, **11**(9):367-372.
114. Yang Z: **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods.** *J Mol Evol* 1994, **39**(3):306-314.
115. Kelchner SA, Thomas MA: **Model use in phylogenetics: nine key questions.** *Trends Ecol Evol (Amst)* 2007, **22**(2):87-94.
116. Sullivan J, Swofford DL: **Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated?** *Syst Biol* 2001, **50**(5):723-729.
117. Waddell PJ, Penny D, Moore T: **Hadamard conjugations and modeling sequence evolution with unequal rates across sites.** *Mol Phylogenet Evol* 1997, **8**(1):33-50.
118. Zwickl DJ, Holder MT: **Model parameterization, prior distributions, and the general time-reversible model in bayesian phylogenetics.** *Syst Biol* 2004, **53**(6):877-888.
119. Beiko RG, Keith JM, Harlow TJ, Ragan MA: **Searching for convergence in phylogenetic Markov Chain Monte Carlo.** *Syst Biol* 2006, **55**(4):553-565.

## **Comprehensive analyses of nuclear rRNA genes to infer Insect phylogeny**

**Sabrina Simon<sup>1,\*</sup> and Heike Hadrys<sup>1,2</sup>**

<sup>1</sup> *ITZ, Ecology & Evolution, Stiftung Tierärztliche Hochschule Hannover, D-30559  
Hannover, Germany*

<sup>2</sup> *Department of Ecology and Evolutionary Biology, Yale University, New Haven,  
Connecticut 06520, USA*

\*Corresponding author

This is the author's version of a work prepared for submission to *BMC Evolutionary Biology*.

## Abstract

**Background:** Phylogenetic studies based on rRNA genes still play a prominent role for resolving relationships across the animal kingdom. Besides focusing on the primary sequence information for the phylogenetic analyses, several studies incorporate the secondary structure of the molecules for different approaches. The secondary structure information is (i) used as alignment guide and supports the process of aligning rRNA sequences across diverse taxa, (ii) essential for the application of biologically realistic mixed DNA/RNA substitution models to the rRNA data and (iii) used a direct source of measurable information to improve phylogenetic reconstructions. Based on sequences of the complete nuclear 28S rRNA we applied these approaches exploring the phylogenetic utility of the 28S rRNA gene at the primary sequence and the secondary structure level to resolve ancient split among the major hexapod lineages.

**Results:** Forty-three new complete 28S rRNA sequences of winged insects (Pterygota) were acquired from partially unrepresented pterygote orders. The updated 28S rRNA data set, in total 123 hexapod species, was analyzed by incorporating specific rRNA evolutionary models and site-specific heterogeneity among the lineages. In addition the phylogenetic analyses were also performed on a concatenated 18S+28S rRNA data set and compared to previously published topologies inferred by these genes. Based on the comparative analyses we show that 28S rRNA sequences contain conflicting phylogenetic signals resulting in misplacements in the inferred topology. Although the concatenation of 28S and 18S rRNA sequences compensates this signal erosion, the data set is still characterized by a low phylogenetic signal for deep splits among hexapods.

**Conclusions:** We demonstrate that despite using sophisticated evolutionary substitution models and incorporating lineage specific substitution rates the resolving power of the nuclear rRNA genes appears limited within the major clades of hexapods. These limitations emphasize the importance of obtaining more data to understand the evolution of macro-evolutionary key innovations among hexapods.

## Background

Ribosomal RNA genes are widely used to reconstruct the phylogeny at family [1], genera [2], order level [3] and even deep phylogenies across the metazoan subgroups [e.g. 4, 5, 6]. Although the multi-gene approaches using protein-coding genes from direct sequencing of target genes, EST (Expressed Sequence Tags) or whole genome projects have provided an increasingly large amount of phylogenetic information to resolve deep metazoan relationships [7-10], nuclear rRNA genes (18S and 28S rRNA) are still invaluable for defining phylogenetic relationships among metazoans. They still provide a much denser taxon coverage due to thousands of sequenced taxa throughout the metazoan kingdom.

The structural feature of ribosomal RNA genes of highly variable sections (expansion segments) between conserved, slowly evolving sections (core regions) makes them to powerful phylogenetic markers which allow studying deep level as well as higher level phylogenies. In addition the conserved core regions are used for the establishment of universal primers to easily and economically amplify the complete genes in overlapping fragments across a broad range of taxa [11]. The length-variable and fast evolving regions, the expansion segments, have been shown to be useful to resolve higher level phylogenies [12, 13], but causes the main difficulties for deep level phylogenetic studies. These variable regions are difficult to align across diverse taxa and are usually omitted in tree reconstruction [14] due to uncertainty of positional homology. Because the topology of a phylogenetic tree may be critically dependent on the accuracy of the sequence alignment employed [15, 16], trees based on rRNA sequences are more likely to reflect the true evolutionary relationships when secondary structure information is available to guide the primary sequence alignment [17]. Several studies have therefore taken into account the secondary structure of the ribosomal RNA genes to increase the dependability of rRNA sequence alignments [5, 18-20] since the secondary structure conservation exceeds primary sequence conservation [21]. The development of new semi-automated tools like MXSCARNA [22] or RNAsalsa [23] further provides a consistent algorithmic framework for the application of secondary structures as alignment guide. These approaches allow the use of the full sequences without omitting the highly variable and hardly align able regions for tree reconstruction and even more important the application of realistic mixed DNA/RNA substitution models for the tree reconstruction. Ribosomal RNA (rRNA) folds on itself, creating a complex secondary structure maintained by bonded nucleotide pairs, which is important for ribosome function [24]. In these stem regions, a substitution at one site is often accompanied by a compensatory substitution at its complement in order to

prevent disruption to the secondary structure of the molecule [25, 26]. Studies demonstrated that it is more appropriate to analyze the alignment in data partitions using a paired-site model for stem regions that treats the nucleotide and its complement as a single character and using standard substitution models for RNA loop regions, which are unpaired [27]. In this respect, it has been demonstrated that the interdependence of character variation among the paired positions (stems) has to be taken into account in order to avoid overemphasizing changes in paired positions [19, 28-30]. Neglecting co-evolving paired sites in stems affects the estimation of bootstrap values and also influences topological inference [12, 29-33].

By taking the secondary structure into account for the establishment of the multiple sequence alignment and the application of mixed DNA/RNA substitution models for tree reconstruction the secondary structure is only indirectly used. However the secondary structure could also be used as a direct source of measurable information [34]. The molecular morphology can be used to count the presence and absence of stems and loops and could directly be converted into a binary character matrix for tree reconstruction [35]. Billoud et al. [34] improved this approach and developed a method termed “molecular morphometrics” which uses the geometrical features, bond energies, base composition, etc. as specific characters to construct a phylogenetic tree. Another approach uses the differences in length and extension of the stems and loops which can be either used as distance characters in a distance-based phylogenetic reconstruction. The distance-based approach of molecular morphology has been widely used across several insect clades. Letsch and co-workers [21] demonstrated that the secondary structure of the nearly complete 28S rRNA correctly recovers all deep splits across dragonflies (Anisoptera). Page et al. [36] analyzed structure variation of 12S rRNA within several louse lineages, Niehuis et al. [37] analyzed structure variation of 18S rRNA across burnet moths and Misof and Fleck [38] analyzed 16S rRNA structure variation across odonates. However comparative structure analyses across higher level clades are until missing.

We therefore used nuclear rRNA sequences of hexapods as model data. The choice of the superclass Hexapoda was prompted by: (i) the availability of several complete 28S rRNA sequences and 18S rRNA sequences, (ii) established primer sequences for the amplification of the complete 28S rRNA gene and (iii) numerous phylogenetic problems posed by this diverse class, still unresolved by previous molecular analyses.



The superclass Hexapoda is one of the most studied animal classes. This fact is not surprising regarding that hexapods comprise the most diverse animal group on Earth, representing approximately 60% of the diversity of life [39]. Tracing their history and assessing their interordinal phylogenetic relationships is critical for understanding the macro-evolutionary key innovations which occurred during their evolution. But still, at present unraveling their evolutionary history is one of the major challenges for phylogenetic studies. Despite progress in the resolution of the basal pterygote relationships [8] and the earliest branching extant holometabolan lineages [10, 40] by applying a phylogenomic approach, there is still no consensus at several places of the evolution of insects. Especially evidence on the relationships among the principal lineages of Neoptera (insects with wing flexion) is both limited and controversial. The neopteran lineages diverged probably in the Late Carboniferous (320 MYA (million years ago)) and fossil records indicate that the radiation into a vast number of lineages and species was nearly simultaneous [39, 41]. This circumstance leads to short internodes separating the lineages and affects the ability to resolve their phylogeny based on molecular data due to loss of phylogenetic signal over time, reviewed in [42]. In an attempt to overcome the lack of phylogenetic signal in the data, multi-gene analyses have become the state-of-the-art approach. However, regarding insect phylogeny the vast majority of genome- or EST-projects are based on derived holometabolan species (approx. 70%) and some insect orders are still completely missing.

However, nuclear rRNA genes are available over a broad range of insect lineages. In addition the promising new approaches for (i) the establishment of positional homology and (ii) the application of biological realistic substitution models for the phylogenetic analyses aid in achieving a more robust reconstruction of insect relationships. To infer insect relationships mainly 18S rRNA genes and fragments of 28S rRNA genes have been used, e.g. [19, 43-46]. Recently, von Reumont et al. [5] broadened the data set by adding 103 new nearly complete 18S and 28S rRNA across a variety of arthropod taxa, focusing on the major arthropod relationships. However, regarding the represented insects lineages there are still some orders missing or underrepresented, e.g. Embioptera, Plecoptera, Blattodea, Rhaphidioptera, Planipennia. Here we add several new complete 28S rRNA sequences across a broad range of pterygote orders and focus only on the interordinal relationships of insects.

The purpose is to explore the potential of the complete 28S rRNA gene at the nucleotide level and at the secondary structure level to resolve deep splits within insects. To further allow comparison with concatenated 18S+28S rRNA analyses, 18S rRNA was also included in the analyses. The used data set currently represents the most comprehensive ones regarding the nuclear rRNA characters (complete 18S and 28S rRNA) and represented hexapod taxa (123 taxa).

## **Results and Discussion**

### ***Phylogenetic analyses***

We added 43 new nearly complete 28S rRNA sequences across 20 pterygote orders by using the universal pterygote 28S rRNA primers given in Additional file 1. To conduct the phylogenetic analyses on a concatenated data set (18S+28S), sequences for the complete 18S rRNA were retrieved from GenBank, whereby closely related species as possible were chosen (chimeran concatenated sequences are marked with an asterisk in Table 1). The prealignment (alignment using Muscle [47, 48]) of 18S rRNA sequences comprised 3,031 positions and 3,315 positions after the progressive multiple alignment method, namely RNAsalsa [23]. ALIScore [49] identified 50.47% positions as randomly similar and after exclusion of these randomly positions with Alicut (<http://www.utilities.zfmk.de>) the resulting 18S rRNA alignment comprised 1,642 positions. The 28S rRNA sequence prealignment comprised 6,330 positions and 6,821 positions after RNAsalsa. Of these 63.78% were identified as randomly similar and the final alignment comprised 2,470 positions after Alicut.

In total 123 hexapod species were included in the phylogenetic analyses to investigate the potential of the complete 28S rRNA gene for resolving deep insect relationships (analyzed species given in Table 1). In addition we performed the phylogenetic analyses on a concatenated data set (18S+28S) and compared the topology with the 28S rRNA topology.

**Table 1: Taxa list**

Taxa list of sampled sequences. \* indicates concatenated 18S and 28S rRNA sequences from different species.

Order	family	species	GenBank 28Sno.	Genbank18S	(* = chimara)
Collembola	Sminthuridae	<i>Allacma fusca</i>	EU376054	EU368610	
	Entomobryoidea	<i>Pogonognathellus flavescens</i>	EU376053	EU368607	
	Sminthuridae	<i>Sminthurus viridis</i>	EF199973	EU368609	
	Poduroidea	<i>Tetradontophora bielensis</i>	EU376051	AY555519	
	Entomobryoidea	<i>Cryptopygus antarcticus</i>	EF199971	EU368605	
	Dicyrtomidae	<i>Dicyrtomina saundersi</i>	EF199974	EU368611	
	Poduroidea	<i>Triacanthella</i> sp.	AY859609	AY859610	
	Poduroidea	<i>Gomphiocephalus hodgsoni</i>	EF199969	EU368601	
	Isotomidae	<i>Isotoma viridis</i>	EU376052	AY596361	
	Neelidae	<i>Megalothorax minimus</i>	EF199975	EU368608	
	Entomobryoidea	<i>Orchesella villosa</i>	EF199972	EU368606	
	Poduroidea	<i>Podura aquatica</i>	EF199970	EU368604	
	Diplura	<i>Campodea augens</i>	EF199977	EU368599	
	Japygidae	<i>Catajapyx aquilonaris</i>	EF199978	EU368600	
Protura	Campodeidae	<i>Campodeidae</i> sp.	AY859560	AY859561	
	Campodeidae	<i>Lepidocampa weberi</i>	EU376050	AY037167	
	Acerentomidae	<i>Acerentomon franzi</i>	EF199976	EU368597	
	Berberentomidae	<i>Baculentulus densus</i>	EU376049	AY037169	*
Archaeognatha	Eosentomidae	<i>Eosentomon</i> sp.	EU376047	EU368598	
	Machilidae	<i>Lepismachilis y-signata</i>	EF199980	EU368613	
	Machilidae	<i>Machilis hrabei</i>	EF199981	EU368612	
Thysanura	Machilidae	<i>Pedetontus okajimae</i>	EU376055	EU368614	
	Lepismatidae	<i>Ctenolepisma longicaudata</i>	AY210810	EU368616	
Odonata	Lepismatidae	<i>Lepisma saccharina</i>	EU376048	EU368615	
	Aeshnidae	<i>Brachytron pratense</i>	EU424323	AF461232	
	Gomphidae	<i>Progomphus obscurus</i>	EU424329	AY749909	
	Libellulidae	<i>Sympetrum danae</i>	EU424330	AF461243	
	Lestidae	<i>Chalcolestes viridis</i>	EU424331	AJ421949	
	Aeshnidae	<i>Aeshna juncea</i>	EU424324	AF461231	
	Aeshnidae	<i>Boyeria irene</i>	this study	EU055162	*
	Corduliidae	<i>Oxygastra curtisii</i>	EU424325	DQ008194	
	Corduliidae	<i>Cordulia aenea</i>	EU424326	AF461236	
	Corduliidae	<i>Somatochlora flavomaculata</i>	EU424327	AF461242	
Ephemeroptera	Protoneuridae	<i>Elatoneura glauca</i>	this study	AJ746315	*
	Epiophlebiidae	<i>Epiophlebia superstes</i>	EU424328	AF461247	
	Libellulidae	<i>Leucorrhinia</i> sp.	AY859583	AY859584	
	Heptageniidae	<i>Epeorus sylvicola</i>	EU414715	AY749837	*
	Heptageniidae	<i>Ecdyonurus</i> sp.	this study	AY121137	*
	Baetidae	<i>Baetis</i> sp.	this study	AY749848	*
	Siphonuridae	<i>Siphonura aestivalis</i>	EU414716	DQ008181	*
	Baetidae	<i>Callibaetis ferrugineus</i>	AY859557	AF370791	
	Blaberidae	<i>Blaberus fusca</i>	this study	DQ874112	*
	Blatellidae	<i>Ectobius lapponicus</i>	EU426877	DQ874125	
Blattodea	Polyphagidae	<i>Cryptocercus kyebangensis</i>	this study	EU253777	*
	Blatellidae	<i>Blattella germanica</i>	AF005243	AF005243	
	Blaberidae	<i>Gromphadorhina laevigata</i>	AY210819	AY210820	
	Caelifera	<i>Anacridium aegyptium</i>	EU414723	AY379759	*
	Acrididae	<i>Euprepocnemis plorans</i>	this study	AY626910	*
	Acrididae	<i>Acrida turrita</i>	this study	Z97560	
Ensifera	Acrididae	<i>Sphingonotus</i> sp.	this study	AF370793	*
	Acrididae	<i>Gomphocerinae</i> sp.	AY859546	AY859547	
	Gryllidae	<i>Leptophyes punctatissima</i>	EU414721	AY521867	*
	Gryllidae	<i>Pholidoptera griseoaptera</i>	EU414722	Z97587	*
	Gryllidae	<i>Acheta domesticus</i>	AY859544	X95741	
Dermaptera	Forficulidae	<i>Forficula auricularia</i>	EU426876	Z97594	
	Forficulidae	<i>Apterygida media</i>	this study	AY521837	*
Plecoptera	Perlidae	<i>Isoperla</i> sp.	EU414717	AF461256	*
	Nemouridae	<i>Nemoura cinerea</i>	this study	AF461257	
	Perlidae	<i>Perlodes dispar</i>	this study	EF622774	
	Nemouridae	<i>Nemoura flexuosa</i>	EU414718	Z97595	*
Mantodea	Pteronarcyidae	<i>Pteronarcys reticulata</i>	this study	AY521880	*
	Pteronarcyidae	<i>Pteronarcys sachalina</i>	this study	EF622817	
	Mantidae	<i>Rhombodera</i> sp.	this study	EF383482	*
	Mantidae	<i>Hierodula membranacea</i>	EU414720	AY491194	*
Mantophasmatodea	Mantidae	<i>Mantis religiosa</i>	AY859585	EF363231	
	Mantophasmatidae	<i>Mantophasma zephyra</i>	EU414719	DQ874153	*
	Mantophasmatodea incertae sedis	<i>Tyrannophasma gladiator</i>	EU426875	AY521863	
Phasmatodea	Phasmidae	<i>Carausius morosus</i>	EU426878	X89488	
	Phyllidae	<i>Bacillus rossius</i>	EU426879	AY121180	
Embiidina	Oligotomidae	<i>Haploembia solieri</i>	this study	Z97593	*
Grylloblattodea	Grylloblattidae	<i>Galloisiana yuasai</i> Asahina	this study	DQ457281	
Isoptera	Mastotermitidae	<i>Mastotermes darwiniensis</i>	this study	AY121141	
Homoptera	Termopsidae	<i>Zootermopsis angusticollis</i>	AY859614	AY859615	
	Cercopidae	<i>Cercopis vulnerata</i>	EU414724	AY744798	*
	Cicadellidae	<i>Graphocephala fennahi</i>	this study	DQ532501	*
	Clastopteridae	<i>Clastoptera obtusa</i>	AF304569	AY744784	
Heteroptera	Machaerotidae	<i>Pectinariophyes reticulata</i>	AF304570	AY744778	
	Acanthocomatidae	<i>Cyphostethus tristriatus</i>	this study	AY252322	*
	Pyrrhocoridae	<i>Pyrrhocoris apterus</i>	EU414725	AY627318	*
	Pentatomidae	<i>Rhaphigaster nebulosa</i>	EU426880	X89495	
	Notonectidae	<i>Notonecta glauca</i>	this study	AY252216	*
	Miridae	<i>Harpocera thoracica</i>	EU414726	AY252388	*

Order	family	species	GenBankK 28Sno.	Genbank18S	(* = chimara)
Coleoptera	Carabidae	<i>Abax parallelus</i>	this study	FJ173132	*
	Dermeistidae	<i>Dermeistes maculatus</i>	this study	EF213892	*
	Carabidae	<i>Molops piceus</i>	this study	FJ173125	*
	Silphidae	<i>Silpha obscura</i>	EU426881	AJ810737	
	Tenebrionidae	<i>Pimella</i> sp.	this study	EF363008	*
	Scarabaeidae	<i>Melolontha melolontha</i>	this study	EF487702	
	Tenebrionidae	<i>Tenebrio</i> sp.	AY210843	X07801	
Hymenoptera	Apidae	<i>Nomada</i> sp.	EU414727	AY703484	*
	Apidae	<i>Bombus terrestris</i>	this study	AY773344	*
	Scoliidae	<i>Scolia</i> sp.	EU414728	EF012932	*
	Tenthredinidae	gen. sp.	EU414729	AF423781	*
	Cynipidae	gen. sp.	this study	AF395149	*
	Formicidae	<i>Myrmecia crosland</i>	AB052895	AB121786	
	Vespoidea	<i>Vespa pensylvanica</i>	AY859612	AY859613	
Lepidoptera	Pieridae	<i>Anthocharis cardamines</i>	this study	AF423785	*
	Pieridae	<i>Pieris napi</i>	EU414731	AF423785	*
	Pieridae	<i>Pieris rapae</i>	this study	AF423785	*
	Pyralidae	<i>Eurhypha hortulana</i>	this study	AF286298	*
	Hesperiidae	<i>Thymelicus sylvestris</i> PODA	this study	EU057177	*
	Boreidae	<i>Boreus hyemalis</i>	EU426882	AF423882	
	Panorpidae	<i>Panorpa</i> sp.	this study	DQ008178	*
Mecoptera	Meropidae	<i>Merope tuber</i>	DQ202351	AF286287	
	Chrysopidae	gen. sp.	this study	X89482	*
	Myrmeleonidae	<i>Euroleon nostras</i>	this study	AF423789	*
Planipennia	Mantispidae	<i>Mantispa mandarina</i> NAVAS	this study	EU797400	*
	Myrmeleontidae	gen. sp.	this study	AF012527	*
	Rhaphidiidae	<i>Phaestigma</i> sp.	this study	X89494	*
Rhaphidoptera	Pulicidae	<i>Ctenocephalides felis</i>	EU414732	AF423914	*
Siphonaptera	Leptoceridae	<i>Triaenodes</i> sp.	EU414730	AF286300	*
Trichoptera	Glossosomatidae	gen. sp.	this study	AF286300	*
	Goeridae	<i>Silo</i> sp.	this study	AF286292	*
	Hydroptilidae	<i>Oxyethira rossi</i>	DQ202352	AF423801	*
Diptera	Calliphoridae	<i>Pollenia rudis</i>	this study	AF322425	*
	Bombyliidae	<i>Bombylius major</i>	this study	EF650090	
	Tipulidae	<i>Ctenophora</i> sp.	this study	AY521834	*
	Culicoidea; Culicidae	<i>Anopheles albimanus</i>	L78065.1	L78065	
	Drosophilidae	<i>Drosophila melanogaster</i>	M21017.1	M21017	
	Chironomidae	<i>Chironomus tentans</i>	X99212.1	X99212	
	Simuliidae	<i>Simulium sanctipauli</i>	AF403805	AF403799	
	Culicidae	<i>Aedes albopictus</i>	L22060	X57172	
	Chironomidae	<i>Acrictopus lucens</i>	AJ586562	AJ586561	

Recently, von Reumont et al. [5] showed in their study on arthropod phylogeny that base frequencies significantly differed across taxa in 18S rRNA as well as 28S rRNA data sets. They identified in both genes the Diptera (high A/T content) and Diplura (low A/T content) sequences responsible for the inhomogeneous data set and fitted three sub-models to each data partition to account for the lineage specific substitution patterns. In addition, their findings showed that the time-heterogeneous approach produces improved likelihood values with improved branch lengths estimates and more realistic topology estimates.

We therefore decided to conduct (i) the time-heterogeneous approach with three submodels for 28S and the combined data set (18S+28S) and (ii) the application of mixed DNA/ RNA substitution models in a Bayesian approach using the *PHASE-2.0* software package [50]. Of the two combinations of mixed DNA/RNA models (TN93 +  $\Gamma$  & RNA16J +  $\Gamma$  and HKY85 +  $\Gamma$  & RNA16K +  $\Gamma$ ), the applied Bayes Factor Test ([51, 52], BFT) favored the TN93 +  $\Gamma$  & RNA16J +  $\Gamma$  model combination for both data sets (28S:  $2\ln B_{10} = 164.68$ , harmonic mean  $\ln L_0$  (TN93 +  $\Gamma$  & RNA16J +  $\Gamma$ ) = 41931.86; harmonic mean  $\ln L_1$  (HKY85 +  $\Gamma$  & RNA16K +  $\Gamma$ ) = 42014.20; 18S+28S:  $2\ln B_{10} = 56.34$ , harmonic mean  $\ln L_0$  (TN93 +  $\Gamma$  & RNA16J +  $\Gamma$ ) = 66908.75; harmonic mean  $\ln L_1$  (HKY85 +  $\Gamma$  & RNA16K +

$\Gamma$ ) = 66936.92). Finally all chains which passed a threshold value in a BFT were included for both data sets. The assembled meta-chains included 42 million generations for the 28S data set and 28 million generations for the concatenated 18S+28S data set, burn-ins were discarded. Each resulting consensus tree was rooted with monophyletic Entognatha (Collembola, Protura, Diplura). Node support values for clades were deduced from 30,000 sampled trees for the 28S data set (Figure 1) and from 20,000 sampled trees for the concatenated 18S+28S data set (Figure 2); detailed support values are given in Additional file 2 and Additional file 3.

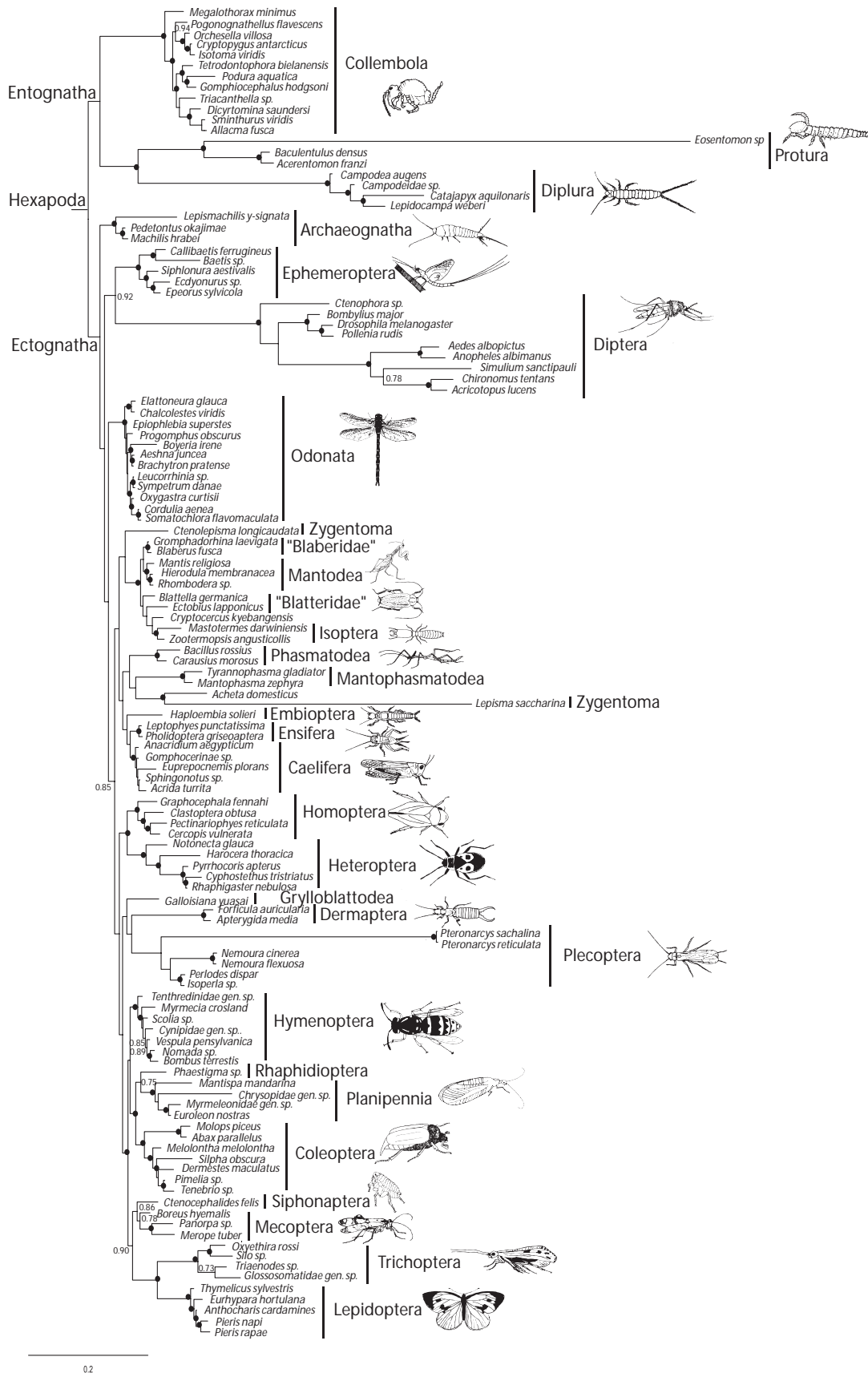
#### *28S resulting topology*

The resulting topology showed several unorthodox positions which are clearly incongruent with morphological evidence and results obtained from other genes.

Both representatives of Zygentoma assumed unorthodox positions in the 28S rRNA tree: *Ctenolepisma longicaudata* formed the sister group of the Dictyoptera clade (Blattodea, Mantodea, Isoptera), and *Lepisma saccharina* clustered with the cricket *Acheta domesticus* in the neopteran assemblage. Consequently, Pterygota always appeared polyphyletic in the analyses based on the complete 28S rRNA sequences. We consider these results highly unlikely, since they contradict all independent evidence from morphology and development.

Another striking feature that is observed in the 28S rRNA topology (Figure 1) is the unorthodox position of Diptera. Diptera formed the sister group of Ephemeroptera and branched at the base of the pterygotes. We assume that the dipteran taxa may be misplaced because of LBA (long-branch attraction) which is also observed in the neighbor-net graph (Additional file 4). The neighbor-net graph which results from a split decomposition based on uncorrected  $p$ -distances showed that the dipteran taxa have long stem lineages, which means that the species share distinct nucleotide patterns not present in other taxa. Consequently the dipteran may be misplaced due to signal erosion or occurrence of homoplasies.

In addition, the resulting topology is poorly supported in many parts, especially the deep nodes within the Hexapod tree. While the intraordinal relationships were well supported ( $pP > 0.95$ ) in most orders, only the clade Holometabola (except for Diptera) and the clade Hemiptera (Homoptera+Heteroptera) received moderate support ( $pP > 0.95$ ) among the interordinal relationships. The overall low support values for the deep nodes reflect the low phylogenetic signal with the complete 28S rRNA gene and indicate a high noise content in the data set due to an excessive rate heterogeneity among the lineages.

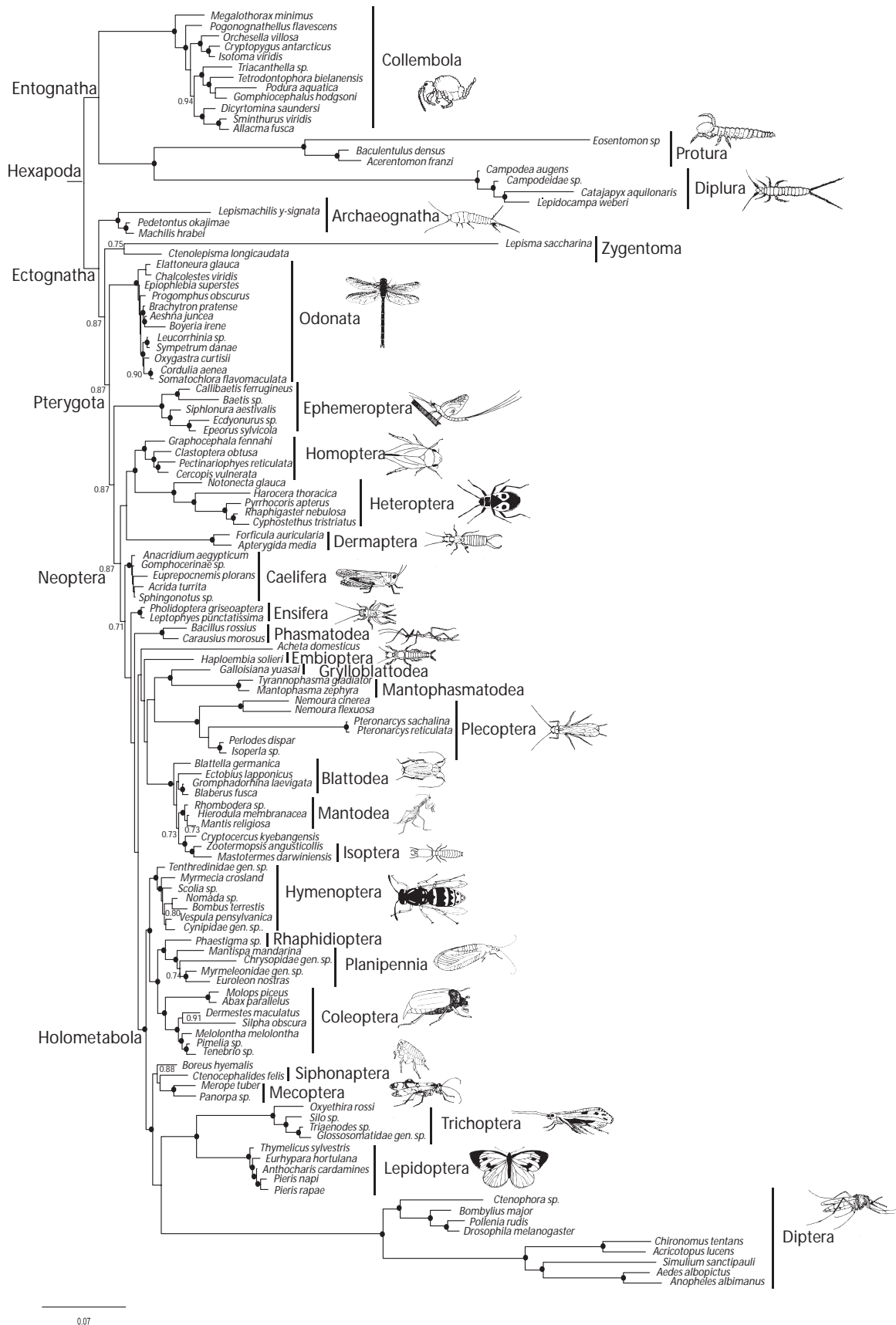
**Figure 1: Time-heterogeneous consensus tree of 28S**

Consensus tree based on the 28S rRNA data set from 30,000 sampled trees of the time-heterogeneous substitution process inferred by *PHASE-2.0*. Support values below 0.70 are not shown (nodes without dots), nodes with a posterior probability (pP>0.95) are represented by dots only.

*Concatenated 18S+28S resulting topology*

In contrast the topology inferred from the concatenated 18S+28S rRNA data set recovers several clades which are supported by morphological characters and other genes (Figure 2).

The consensus tree showed the expected paraphyly of primarily wingless insects with Archaeognatha as a sister group to Zygentoma + Pterygota. Odonates emerged first within Pterygota and formed the sister group to a monophyletic clade of Ephemeroptera + Neoptera, supporting the Chiasmomyaria hypothesis [53]. Within the Neoptera, Dermaptera clustered with the monophyletic Hemiptera (Homoptera+Heteroptera). Orthoptera (Caelifera+Ensifera) appeared paraphyletic and even polyphyletic when *Acheta* was included. However, a position of *Acheta domesticus* outside an orthopteran assemblage or even ensiferan assemblage can be observed in every phylogenetic study using the 28S rRNA sequence, e.g. [4, 5, 54]. Galloisiana (Grylloblattodea) emerged with monophyletic Mantophasmatodea and this assemblage formed the sister group to Plecoptera. Blattodea were paraphyletic with respect to the sister group relationship of Cryptocercus with the isopteran representatives. Dictyoptera (Blattodea, Mantodea, Isoptera) were recovered as a monophyletic clade with high support (pP 1.0). Within Holometabola, Hymenoptera emerged as a sister group to the clade of Coleoptera and Neuropterida (Planipennia + Rhaphidioptera). Mecoptera appeared paraphyletic with Boreidae as sister group to the clade formed by Siphonaptera and the remaining Mecopteran. While the paraphyly of Mecoptera is supported by different genes, Boreidae is assumed to form the sister group to Siphonaptera [55]. While we could recover Amphiesmenoptera (Trichoptera + Lepidoptera), we failed to recover Antliophora (Diptera(Mecoptera+Siphonaptera)) with our analyses.

**Figure 2: Time-heterogeneous consensus tree of concatenated 18S+28S**

Consensus tree based on the concatenated 18+28S rRNA data set from 20,000 sampled trees of the time-heterogeneous substitution process inferred by *PHASE-2.0*. Support values below 0.70 are not shown (nodes without dots), nodes with a posterior probability (pP>0.95) are represented by dots only.



Based on the phylogenetic analyses and the inferred topologies we could show the major problems when estimating relationships among the basal lineages of Hexapoda. Although (i) we have incorporated biologically realistic model parameters, such as site interaction (mixed DNA/RNA models) and compositional heterogeneity of base frequency, (ii) discarded randomly similar aligned positions to reduce the signal-to-noise ratio and (iii) tried to break down long branches by a dense taxon sampling, we still have several misplacements in the 28S rRNA tree as well as in the 18S+28S rRNA tree. These misplacements are more striking among the principal lineages of Neoptera. The shortness of the internodes of the neopteran lineages in the trees reflects the divergence in the ancient past with a nearly simultaneously radiation into a vast number of species. While adding taxa certainly helped with the resolution within orders, some long-branch problems persisted. We cannot clearly address the reason but assume that saturation by multiple substitutions caused signal erosion which yielded to low support values.

It is particularly noticeable how the inferred insect relationships are prone to taxon sampling. Through comparison of the inferred topologies in the study of von Reumont et al. [5] and in this study several incongruences among the insect relationships are observed. Although both studies support the monophyly of Ectognatha, Pterygota and Neoptera the support values generally decreases by extending taxon sampling. Moreover, both studies support the monophyly of Holometabola, but the inferred earliest branching of extant holometabolan lineages are incongruent. Von Reumont et al. [5] support hymenopterans as the sister group to all other holometabolan insects and corroborate previous studies [e.g. 8, 10, 40]. However, we could not confirm this earliest branching pattern in the holometabolan assemblage. Particular attention should be paid to the incongruence among the inferred topologies regarding the basal neopteran lineages (Polyneoptera). There is overall little support for any relationship among polyneopteran orders, except for the Dictyoptera clade (Blattodea(Mantodea(Cryptocercus+Isoptera))). In contrast, based on the presented analyses the inferred basal pterygote relationships are congruent to the results of von Reumont et al. [5]. Both studies show strong support for Odonata as the earliest branching pterygote lineage and Ephemeroptera as sister group to the neopteran lineages (Chiaatomyaria).

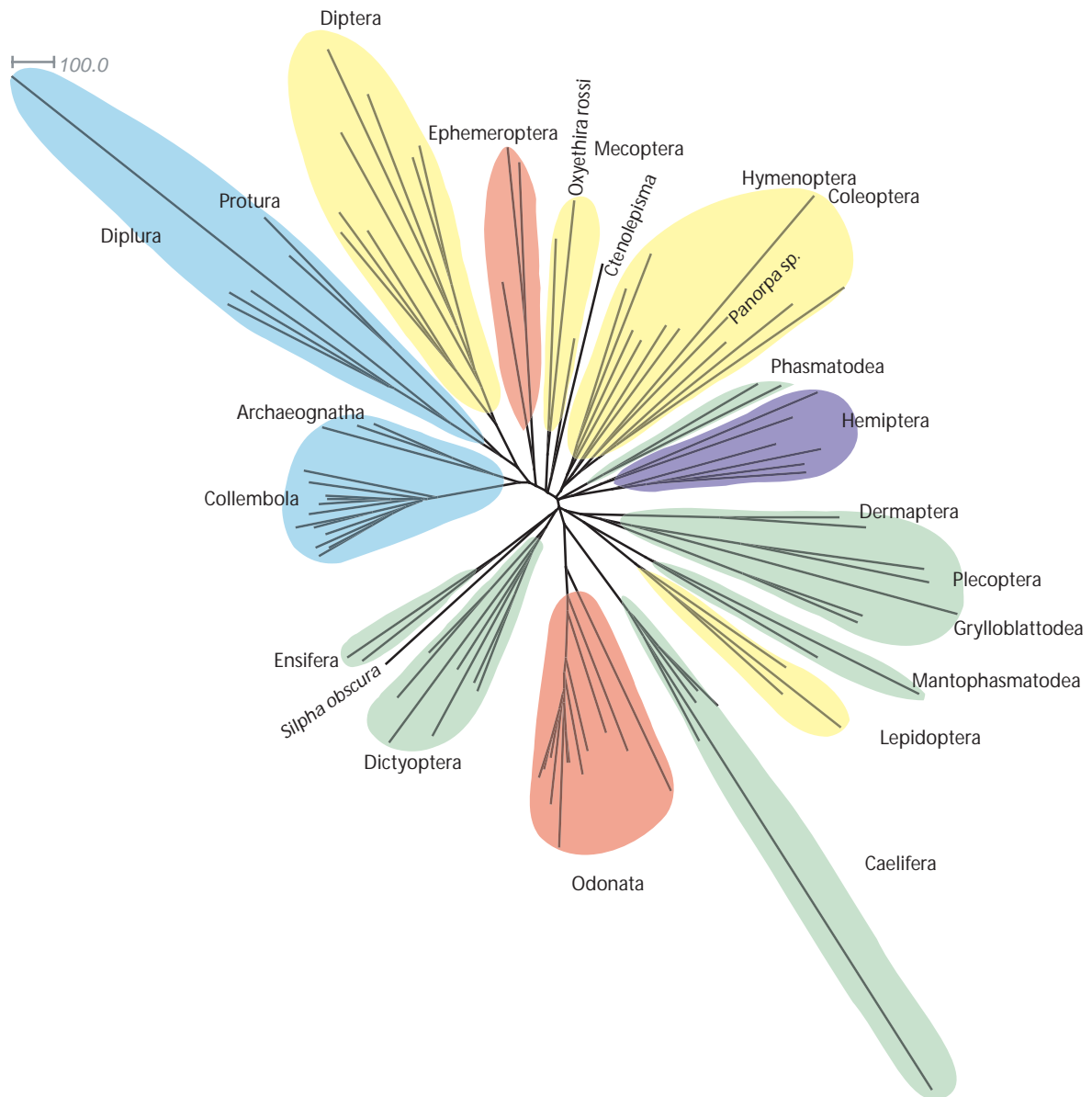
### ***Structure analysis of full length of 28S rRNA***

We further examined the phylogenetic signal of structure motifs of the complete 28S rRNA molecule to uncover deep insect relationships and applied a quantitative approach. The NJ tree reconstruction, which results from edit distances of “constrained full MFE structures”, reflects the overall low phylogenetic signal for deep splits among insect lineages. The inferred topology is in most places incongruent to the topology based on the primary sequence (Figure 3). Although the topology supports the monophyly of most hexapod orders (except for Coleoptera, Mecoptera, Hymenoptera), the relationships between the orders is not congruent to any hypothesized relationship. The recovered monophyly of most of the orders could indirectly indicate correct structure inference, but the lack of resolution for the interordinal relationships uncovers insufficient phylogenetic signal in structure variation for deep phylogenetic splits.

In addition some placements indicate the existence of problems for deep phylogenetic studies. Some taxa such as *Ctenolepisma longicaudata* (clearly separated from the other “apterygote” taxa), *Silpha obscura* (which do not cluster with the other coleopteran species), *Oxyethira rossi* (clusters with mecopteran Merope), *Panorpa* sp.(clusters with Hymenoptera and Coleoptera) are misplaced in the topology. This could be the effect of incorrect structure inference for these taxa or the occurrence of homoplasies. After exclusion of the holometabolous orders (Mecoptera, Trichoptera, Coleoptera, Lepidoptera and Hymenoptera) the zgyentoma *Ctenolepisma* clusters with the Archaeognatha (see Additional file 5) which is more congruent to previously published hypotheses and indicate the correct structure inference for this species.

b) The clade comprising Hymenoptera, Coleoptera and the mecopteran *Panorpa* do not recover the accepted monophyly of Hymenoptera and Coleoptera. Consequently these taxa may be also misplaced due to signal erosion or occurrence of homoplasies.

Recently Letsch et al. [21] showed that the structure variation in the 28S rRNA gene contains phylogenetic signal to recover expected deep splits in Anisoptera (dragonfly) evolution. However, our results demonstrate the existing lack of phylogenetic information in structure variation for ancient splits among hexapods.



**Figure 3: Geometric distances of the complete 28S secondary structures.**

Complete structures. Geometric distances of the complete 28S secondary structures, presented by a NJ tree. Primarily wing-less insects are illustrated in blue, “Palaeoptera” (Ephemeroptera&Odonata) in red, Polyneoptera in green, Paraneoptera in violet and Holometabola in yellow.

## Conclusions

The incorporation of additional data, such as structural information to identify homologous positions between sequences, is a promising approach to improve alignments and phylogenetic analyses by applying mixed DNA/RNA substitution models. Within the Hexapoda, where the sequence composition is highly variable between orders, both alignment methods and models of evolution used in phylogenetic analyses still remain the main bottlenecks in tracing deep phylogeny.

Semi-automated tools like RNAsalsa [23] provide a consistent algorithmic framework for the application of secondary structures as alignment guides and for the application of realistic mixed DNA/RNA substitution models for tree reconstruction. Moreover, the secondary structure analyses of full length 28S rRNA indicates correct structure inference of RNAsalsa through comparative and thermodynamic folding.

However, despite the progress of bioinformatic methods for rRNA data the resolution of deep splits among insects remains difficult. Based on our results we show that the 28S rRNA is characterized by a signal-to-noise ratio which makes phylogenetic analyses problematic and results in several misplacements even if realistic time-heterogeneous and mixed models (DNA/RNA) for phylogenetic reconstructions were applied. Although the combined approach using 18S and 28S rRNA sequences compensates the signal erosion in several cases, an overall low support for the short internodes connecting the insect lineages is observed. In addition we have shown the extreme sensitivity of taxon sampling for the reconstruction of ancient splits in insects. We agree with von Reumont et al. [5] that rRNA data cannot robustly resolve the most ancient splits within insects. Moreover, we showed the sensitivity of rRNA based insect phylogenies on taxon sampling. Nonetheless, some inferred relationships remain stable indicating a robust reconstructed phylogeny, e.g. the Chiasmomyaria hypothesis (Odonata, Ephemeroptera+Neoptera).

We conclude that the future of hexapod class level phylogeny appears not to lie in rRNA genes. However, the addition of data (molecular data and/or morphology) may help to accurately track these short ancient internodes connecting the taxonomic groups. Whether the accumulation of complete genomes and EST projects for promising multi-gene approaches will be more effective to resolve such ancient splits has to be awaited.

## **Methods**

### **Molecular methods**

DNA was extracted from ethanol (98%) preserved animals according to a modified standard protocol [56]. For larger specimens a small tissue from a single leg or alternatively wing muscle from the mesothorax was dissected. For smaller specimens the entire thorax was used after removing of the gut. The complete 28S rRNA was amplified by using different primer combinations (for primer combinations see [5] and for primer sequences see Supplement Table 1). For specimens where the primer combination 28II-28jj

failed and the primer combinations 28ll-28hh and 28w-28jj were alternatively used. The amplification and sequencing procedure for the 28S rRNA sequences were performed as described in von Reumont et al. [5].

### Multiple sequence alignment

The source of the sequence data (18S and 28S) employed for the analyses is listed in Table 1. The 28S rRNA and 18S rRNA sequences were separately aligned with Muscle [48] under default parameters and manually modified for species with missing data.

Afterwards RNAsalsa [23] was applied for both alignments using default parameters and using the 18S/28S structure of *Anopheles albimanus* as a constraint to simultaneously align sequence and resulting structure strings. Both alignments were investigated for randomly similar positions using ALISCORE [49] with a window size of 4. Gaps were treated as ambiguities (-N option) and the maximal amount of random pairwise comparisons were analyzed. The identified randomly similar positions were excluded using Alicut (<http://www.utilities.zfmk.de>) maintaining stem positions (-r option) even if identified as randomly similar by ALISCORE through inclusion of the resulting consensus structure (18S/28S) of RNAsalsa in the alignment. The final 28S rRNA alignment and the concatenated 18S and 28S alignment were used for the phylogenetic analyses with PHASE-2.0 [50].

### Phylogenetic analyses

All analyses were performed for the 28S rRNA alignment and the concatenated 18S+28S rRNA data set. Mixed RNA/DNA substitution models were chosen, where loop regions were governed by DNA models and stem regions by RNA models that consider co-variation. Compositional heterogeneity of base frequency was evaluated as described elsewhere [5]. In order to take inhomogeneous base composition among taxa into account all phylogenetic analyses were performed in PHASE-2.0. To select the best model set the preruns (28S and concatenated 18S+28S) were performed as described in von Reumont et al [5] and the model with the best fitness was chosen applying a Bayes Factor Test [51, 52] to the positive values of the harmonic means calculated from the  $\ln L$  values. We conducted the heterogeneous approach where exchangeability parameters (average substitution rate ratio values, rate ratios and alpha shape parameter) were fixed as input values. Values for these parameters were estimated as described in von Reumont et al. [5]. To reflect the base

frequency heterogeneity among taxa the number of base frequency groups was set to three submodels.

In total 8 independent chains of 7,000,000 generations for both datasets were run. For each chain the first two million generations were discarded as burn-in (sampling period of 1000). For both datasets all chains which passed a threshold value in a BFT ( $2\ln B_{10}$ -value  $< 10$ ) were assembled to a metachain [57] using a Perl script, kindly provided by O. Niehuis and modified by K. Meusemann. Consensus trees and posterior probability values were inferred using *mcmcsummarize*. Each resulting consensus tree was rooted with monophyletic Entognatha (Collembola, Protura, Diplura).

### ***Structure analysis of full 28S rRNA***

We further investigated the phylogenetic value of the full secondary structure of 28S rRNA to resolve deep splits in insect evolution. For this approach we only included species for which the complete sequence was available (93 taxa). We used the “full constrained MFE structures” which consisted of all predicted base pairings after the constrained thermodynamic folding step provided by RNAsalsa. We further applied a quantitative approach to use the structural differences as phylogenetic characters. Therefore we used a distance-based phylogenetic reconstruction where the differences of secondary structures are represented as tree edit distances  $d_t$  calculated by RNAdistance from the ViennaRNAPackage [58]. This measurement counts the minimum number of insertions and deletions of paired and unpaired bases needed to transform one RNA structure into another [59, 60]. The distance matrix representing all possible number of pairwise comparisons were used to reconstruct the Neighbor-Joining (NJ) tree with MEGA 4.0 [61].

### **List of abbreviations**

rRNA: ribosomal RNA; pP: posterior probability;  $\ln$ : natural logarithm; BFT: Bayes Factor Test, NJ: Neighbor-Joining; MFE: minimum free energy.

### **Author’s contribution**

SS was primarily responsible for the design of the study, conducted the experiments and analyses and drafted the primary version of the manuscript. Both authors discussed and approved the final manuscript.

**Additional material****Additional file 1**

Primer list.

**Additional file 2**

Bayesian support values of 28S rRNA topology for selected clades. List of Bayesian support values (posterior probability, pP) of the inferred 28S rRNA topology for selected clades of the time-heterogeneous.

**Additional file 3**

Bayesian support values of 18S+28S rRNA topology for selected clades. List of Bayesian support values (posterior probability, pP) of the inferred 18S+28S rRNA topology for selected clades of the time-heterogeneous.

**Additional file 4**

Neighbornet graph of the 28S rRNA alignment. Neighbornet graph based on uncorrected *p*-distances constructed in SplitsTree4 using the 28S rRNA alignment after exclusion of randomly similar sections evaluated with ALISCORE.

**Additional file 5**

Geometric distances of the complete 28S secondary structures. Geometric distances of the complete 28S secondary structures with the exclusion of Mecoptera, Trichoptera, Coleoptera, Lepidoptera and Hymenoptera, presented by a NJ tree.

**Acknowledgements**

We thank Christian Epe, Ryuichiro Machida, Albert Melber, Mike D. Picker, Reinhard Predel, Sven Sagasser and Valentina Teslenko for their help in collecting specimens and for providing tissue or other DNA-samples. Thanks also go to Berit Ullrich, Oliver Niehuis and Karen Meusemann for providing Perl-Scripts. This work was supported by the German Science Foundation (DFG) in the priority program SPP 1174 "Deep Metazoan Phylogeny", DFG grant HA 1947/5.

**References**

1. Fitch DH, Bugaj-Gaweda B, Emmons SW: **18S ribosomal RNA gene phylogeny for some Rhabditidae related to Caenorhabditis**. *Mol Biol Evol* 1995, **12**(2):346-358.
2. Jordal B, Gillespie JJ, Cognato AI: **Secondary structure alignment and direct optimization of 28S rDNA sequences provide limited phylogenetic resolution in bark and ambrosia beetles (Curculionidae: Scolytinae)**. *Zool Scr* 2008, **37**(1):43-56.

3. Xie Q, Tian Y, Zheng L, Bu W: **18S rRNA hyper-elongation and the phylogeny of Euhemiptera (Insecta: Hemiptera)**. *Mol Phylogenet Evol* 2008, **47**(2):463-471.
4. Mallatt J, Craig CW, Yoder MJ: **Nearly complete rRNA genes assembled from across the metazoan animals: effects of more taxa, a structure-based alignment, and paired-sites evolutionary models on phylogeny reconstruction**. *Mol Phylogenet Evol* 2010, **55**(1):1-17.
5. von Reumont BM, Meusemann K, Szucsich NU, Dell'Ampio E, Gowri-Shankar V, Bartel D, Simon S, Letsch HO, Stocsits RR, Luan YX *et al*: **Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships**. *BMC Evol Biol* 2009, **9**:119.
6. Passamanek Y, Halanych KM: **Lophotrochozoan phylogeny assessed with LSU and SSU data: evidence of lophophorate polyphyly**. *Mol Phylogenet Evol* 2006, **40**(1):20-28.
7. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD *et al*: **Broad phylogenomic sampling improves resolution of the animal tree of life**. *Nature* 2008, **452**(7188):745-749.
8. Simon S, Strauss S, von Haeseler A, Hadrys H: **A phylogenomic approach to resolve the basal pterygote divergence**. *Mol Biol Evol* 2009, **26**(12):2719-2730.
9. Philippe H, Derelle R, Lopez P, Pick K, Borchellini C, Boury-Esnault N, Vacelet J, Renard E, Houliston E, Queinnec E *et al*: **Phylogenomics revives traditional views on deep animal relationships**. *Curr Biol* 2009, **19**(8):706-712.
10. Savard J, Tautz D, Richards S, Weinstock GM, Gibbs RA, Werren JH, Tettelin H, Lercher MJ: **Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects**. *Genome Res* 2006, **16**(11):1334-1338.
11. Hillis DM, Dixon MT: **Ribosomal DNA: molecular evolution and phylogenetic inference**. *Q Rev Biol* 1991, **66**(4):411-453.
12. Voigt O, Erpenbeck D, Worheide G: **Molecular evolution of rDNA in early diverging Metazoa: first comparative analysis and phylogenetic application of complete SSU rRNA secondary structures in Porifera**. *BMC Evol Biol* 2008, **8**:69.
13. Gillespie JJ, Munro JB, Heraty JM, Yoder MJ, Owen AK, Carmichael AE: **A secondary structural model of the 28S rRNA expansion segments D2 and D3 for chalcidoid wasps (Hymenoptera: Chalcidoidea)**. *Mol Biol Evol* 2005, **22**(7):1593-1608.
14. Wuyts J, De Rijk P, Van de Peer Y, Pison G, Rousseeuw P, De Wachter R: **Comparative analysis of more than 3000 sequences reveals the existence of two pseudoknots in area V4 of eukaryotic small subunit ribosomal RNA**. *Nucleic Acids Res* 2000, **28**(23):4698-4708.
15. Feng DF, Doolittle RF: **Progressive sequence alignment as a prerequisite to correct phylogenetic trees**. *J Mol Evol* 1987, **25**(4):351-360.
16. Olsen GJ, Woese CR: **Ribosomal RNA: a key to phylogeny**. *Faseb J* 1993, **7**(1):113-123.
17. Schnare MN, Damberger SH, Gray MW, Gutell RR: **Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23 S-like) ribosomal RNA**. *J Mol Biol* 1996, **256**(4):701-719.
18. Kjer KM: **Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs**. *Mol Phylogenet Evol* 1995, **4**(3):314-330.



19. Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A: **Towards an 18S phylogeny of hexapods: accounting for group-specific character covariance in optimized mixed nucleotide/doublet models.** *Zoology (Jena)* 2007, **110**(5):409-429.
20. Hickson RE, Simon C, Perrey SW: **The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence.** *Mol Biol Evol* 2000, **17**(4):530-539.
21. Letsch HO, Greve C, Kuck P, Fleck G, Stocsits RR, Misof B: **Simultaneous alignment and folding of 28S rRNA sequences uncovers phylogenetic signal in structure variation.** *Mol Phylogenet Evol* 2009, **53**(3):758-771.
22. Tabei Y, Kiryu H, Kin T, Asai K: **A fast structural multiple alignment method for long RNA sequences.** *BMC Bioinformatics* 2008, **9**:33.
23. Stocsits RR, Letsch H, Hertel J, Misof B, Stadler PF: **Accurate and efficient reconstruction of deep phylogenies from structured RNAs.** *Nucleic Acids Res* 2009, **37**(18):6184-6193.
24. Noller HF: **Structure of ribosomal RNA.** *Annu Rev Biochem* 1984, **53**:119-162.
25. Smith AD, Lui TWH, Tillier ERM: **Empirical models for substitution in ribosomal RNA.** *Molecular Biology and Evolution* 2004, **21**(3):419-427.
26. Tillier ERM, Collins RA: **High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA.** *Genetics* 1998, **148**(4):1993-2002.
27. Ware JL, Ho SY, Kjer K: **Divergence dates of libelluloid dragonflies (Odonata: Anisoptera) estimated from rRNA using paired-site substitution models.** *Mol Phylogenet Evol* 2008, **47**(1):426-432.
28. Schöninger M, von Haeseler A: **A stochastic model for the evolution of autocorrelated DNA sequences.** *Mol Phylogenet Evol* 1994, **3**(3):240-247.
29. Jow H, Hudelot C, Rattray M, Higgs PG: **Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution.** *Mol Biol Evol* 2002, **19**(9):1591-1601.
30. Galtier N: **Sampling properties of the bootstrap support in molecular phylogeny: Influence of nonindependence among sites.** *Syst Biol* 2004, **53**(1):38-46.
31. Tsagkogeorga G, Turon X, Hopcroft RR, Tilak MK, Feldstein T, Shenkar N, Loya Y, Huchon D, Douzery EJ, Delsuc F: **An updated 18S rRNA phylogeny of tunicates based on mixture and secondary structure models.** *BMC Evol Biol* 2009, **9**:187.
32. Telford MJ, Wise MJ, Gowri-Shankar V: **Consideration of RNA secondary structure significantly improves likelihood-based estimates of phylogeny: Examples from the Bilateria.** *Mol Biol Evol* 2005, **22**(4):1129-1136.
33. Dixon MT, Hillis DM: **Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis.** *Mol Biol Evol* 1993, **10**(1):256-267.
34. Billoud B, Guerrucci MA, Masselot M, Deutsch JS: **Cirripede phylogeny using a novel approach: molecular morphometrics.** *Mol Biol Evol* 2000, **17**(10):1435-1445.
35. Ender A, Schierwater B: **Placozoa are not derived cnidarians: evidence from molecular morphology.** *Mol Biol Evol* 2003, **20**(1):130-134.
36. Page RD, Cruickshank R, Johnson KP: **Louse (Insecta: Phthiraptera) mitochondrial 12S rRNA secondary structure is highly variable.** *Insect Mol Biol* 2002, **11**(4):361-369.
37. Niehuis O, Naumann CM, Misof B: **Phylogenetic analysis of Zygaenoidea small-subunit rRNA structural variation implies initial oligophagy on cyanogenic**

- host plants in larvae of the moth genus *Zygaena* (Insecta : Lepidoptera).** *Zool J Linn Soc-Lond* 2006, **147**(3):367-381.
38. Misof B, Fleck G: **Comparative analysis of mt LSU rRNA secondary structures of Odonates: structural variability and phylogenetic signal.** *Insect Mol Biol* 2003, **12**(6):535-547.
  39. Grimaldi DA, Engel MS: **Evolution of the Insects.** New York: Cambridge University Press; 2005.
  40. Wiegmann BM, Trautwein MD, Kim JW, Cassel BK, Bertone MA, Winterton SL, Yeates DK: **Single-copy nuclear genes resolve the phylogeny of the holometabolous insects.** *BMC Biol* 2009, **7**:34.
  41. Labandeira CC, Sepkoski JJ, Jr.: **Insect diversity in the fossil record.** *Science* 1993, **261**(5119):310-315.
  42. Whitfield JB, Kjer KM: **Ancient rapid radiations of insects: challenges for phylogenetic analysis.** *Annu Rev Entomol* 2008, **53**:449-472.
  43. Kjer KM: **Aligned 18S and insect phylogeny.** *Syst Biol* 2004, **53**(3):506-514.
  44. Kjer KM, Carle FL, Litman J, Ware J: **A molecular phylogeny of Hexapoda.** *Arthropod Systematics & Phylogeny* 2006, **65**:35-44.
  45. Wheeler WC, Whiting MF, Wheeler QD, Carpenter JM: **The Phylogeny of the Extant Hexapod Orders.** *Cladistics* 2001, **17**:113-169.
  46. Ogden TH, Whiting MF: **The problem with "the Paleoptera Problem:" sense and sensitivity.** *Cladistics* 2003, **19**:432-442.
  47. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**:113.
  48. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32**(5):1792-1797.
  49. Misof B, Misof K: **A Monte Carlo Approach Successfully Identifies Randomness in Multiple Sequence Alignments : A More Objective Means of Data Exclusion.** *Systematic Biol* 2009, **58**(1):21-34.
  50. Gowri-Shankar V, Jow H: **PHASE: a software package for Phylogenetics And Sequence Evolution. 2.0.** In.: University of Manchester; 2006.
  51. Kaas RE, Raftery AE: **Bayes Factors.** *Journal of the American Statistical Association* 1995, **90**(430):773-795.
  52. Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JeL: **Bayesian phylogenetic analysis of combined data.** *Syst Biol* 2004, **53**(21):47-67.
  53. Boudreaux HB: **Arthropod phylogeny with special reference to insects.** New York: Wiley; 1979.
  54. Mallatt J, Giribet G: **Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch.** *Mol Phylogenet Evol* 2006, **40**(3):772-794.
  55. Whiting MF: **Mecoptera is paraphyletic: multiple genes and phylogeny of Mecoptera and Siphonaptera.** *Zoological Scripta* 2001, **31**:93-104.
  56. Hadrys H, Schierwater B, Dellaporta SL, DeSalle R, Buss LW: **Determination of paternity in dragonflies by Random Amplified Polymorphic DNA fingerprinting.** *Mol Ecol* 1993, **2**(2):79-87.
  57. Beiko RG, Keith JM, Harlow TJ, Ragan MA: **Searching for convergence in phylogenetic Markov Chain Monte Carlo.** *Syst Biol* 2006, **55**(4):553-565.
  58. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P: **Fast Folding and Comparison of Rna Secondary Structures.** *Monatsh Chem* 1994, **125**(2):167-188.
  59. Fontana W, Konings DAM, Stadler PF, Schuster P: **Statistics of Rna Secondary Structures.** *Biopolymers* 1993, **33**(9):1389-1404.

60. Moulton V, Zuker M, Steel M, Pointon R, Penny D: **Metrics on RNA secondary structures**. *J Comput Biol* 2000, **7**(1-2):277-292.
61. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0**. *Mol Biol Evol* 2007, **24**(8):1596-1599.

## **A Phylogenomic Approach to Resolve the Basal Pterygote Divergence**

**Sabrina Simon<sup>1,\*</sup>, Sascha Strauss<sup>2</sup>, Arndt von Haeseler<sup>2</sup>, Heike Hadrys<sup>1,3</sup>**

<sup>1</sup> *ITZ, Ecology & Evolution, Stiftung Tierärztliche Hochschule Hannover, D-30559 Hannover, Germany*

<sup>2</sup> *Center for Integrative Bioinformatics Vienna, Max F Perutz Laboratories, University of Vienna, Medical University of Vienna, University of Veterinary Medicine, 1030 Vienna, Austria.*

<sup>3</sup> *Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06520, USA*

\*Corresponding author

This is the author's version of a work originally published by the Oxford University Press in: *Molecular Biology and Evolution* 2009 Dec;26(12):2719-2730; available under doi:10.1093/molbev/msp191

**Abstract**

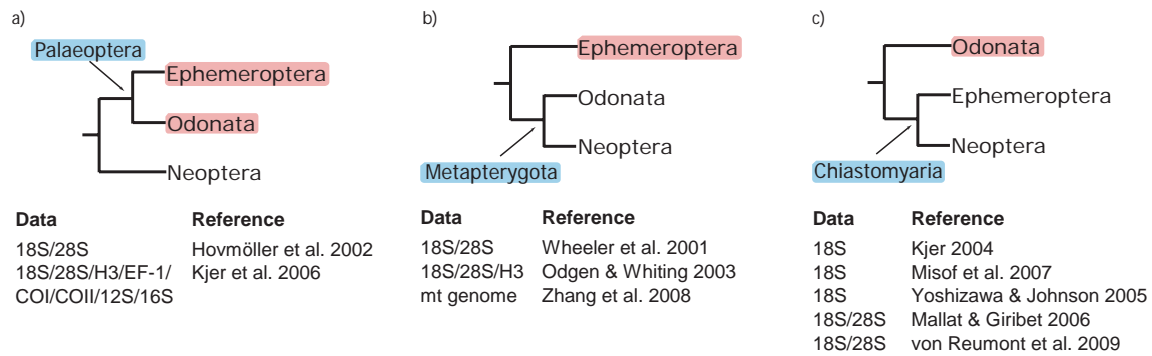
One of the most fascinating Bauplan transitions in the animal kingdom was the invention of insect wings, a change which also contributed to the success and enormous diversity of this animal group. However, the origin of insect flight and the relationships of basal winged insect orders are still controversial. Three hypotheses have been proposed to explain the phylogeny of winged insects: (i) the traditional Palaeoptera hypothesis (Ephemeroptera+Odonata, Neoptera), (ii) the Metapterygota (Ephemeroptera, Odonata+Neoptera) and (iii) the Chiasmomyaria hypothesis (Odonata, Ephemeroptera+Neoptera). Neither phylogenetic analyses of single genes nor even multiple marker systems (e.g. molecular markers + morphological characters) have yet been able to conclusively resolve basal pterygote divergences. A possible explanation for the lack of resolution is that the divergences took place in the mid-Devonian within a short period of time, and attempts to solve this problem have been confounded by the major challenge of finding molecular markers to accurately track these short ancient internodes. Although phylogenomic data are available for Neoptera and some wingless (apterygote) orders, they are lacking for the crucial Odonata and Ephemeroptera orders. We adopt a multi-gene approach including data from two new EST projects – from the orders Ephemeroptera (*Baetis* sp.) and Odonata (*Ischnura elegans*) – to evaluate the potential of phylogenomic analyses in clarifying this unresolved issue. We analyzed two data sets that differed in represented taxa, genes and overall sequence lengths: *maxspe* (15 taxa, 125 genes, 31,643 amino acid positions), *maxgen* (8 taxa, 150 genes, 42,541 amino acid positions). Maximum likelihood and Bayesian inference analyses both place the Odonata at the base of the winged insects. Furthermore, statistical hypotheses testing rejected both the Palaeoptera and the Metapterygota hypotheses. The comprehensive molecular data set developed here provides conclusive support for odonates as the most basal winged insect order (Chiasmomyaria hypothesis). Data quality assessment indicates that proteins involved in cellular processes and signaling harbor the most informative phylogenetic signal.

**Key words:** Basal pterygote divergence, Palaeoptera, Metapterygota, Chiasmomyaria, phylogenomics, expressed sequence tags

## Introduction

Insects are the most diverse animal group on earth and dominate every ecosystem except the benthic zone (Grimaldi and Engel 2005). The winged insects – Pterygota – account for more than 98% off the class Insecta (Grimaldi and Engel 2005). According to fossil records, flying insects originated in the Early Carboniferous period (approx. 320 MYA), whereas a DNA-based study suggested an origin in the mid-Devonian (approx. 387 MYA) (Gaunt and Miles 2002). A recent analysis of Engel and Grimaldi (2004) suggested that the origin of insect wings occurred coincident with the development of arborescence, and agreed with the molecular estimates of Gaunt and Miles (2002). With the invention of the wings, insects were able to invade every ecosystem, escape predators, and exploit scattered resources, resulting in rapid radiations into vast numbers of species (Hennig 1969). Considering the tremendous impact this change produced, the evolution of the flying insects is one of the most fascinating questions in evolutionary biology. Martynov (1925) was the first to distinguish two groups of winged insects based on wing function – Palaeoptera and Neoptera. He assumed the inability to fold back the wings, as seen in Ephemeroptera and Odonata, to be an ancestral condition and therefore called them Palaeoptera (old wings) in contrast to those with this ability, which he called Neoptera (new wings). The monophyly of Palaeoptera has been controversial ever since. In contrast to the accepted monophyly of Neoptera, the so-called “Palaeoptera Problem” is one of the unsolved mysteries in insect systematics.

Today three hypotheses are proposed to explain the phylogenetic relationships of the basal winged insects: (i) the Palaeoptera scenario which supports a basal sister group position of Odonata and Ephemeroptera (Odonata+Ephemeroptera, Neoptera), (ii) the Metapterygota scenario (Ephemeroptera basal, Odonata+Neoptera) and (iii) the Chiasmomyaria scenario (Odonata basal, Ephemeroptera+Neoptera) (Whitfield and Kjer 2008) (Fig.1). Each hypothesis is still considered viable and supported by morphological as well as molecular data. Moreover, some molecular data using the same genes support all three hypotheses depending on the analyses applied (e.g. Hovmöller et al. 2002; Ogden and Whiting 2003; Mallatt and Giribet 2006).



**Fig. 1.** – The three hypotheses at the base of the pterygotes: (a) Palaeoptera (Ephemeroptera+Odonata, Neoptera), (b) Metapterygota (Ephemeroptera, Odonata+Neoptera), (c) Chiasmomyaria (Odonata, Ephemeroptera+Neoptera). The sister group relationships are indicated in blue and the resulting basal pterygote order in red. Below are different molecular studies listed supporting one of the three hypotheses partly using the same genes.

The Palaeoptera are a morphologically well-supported group due to the fact that the Odonata and Ephemeroptera are unable to flex their wings back over the abdomen while members of the Neoptera harbor the necessary muscles and wing sclerites for this movement (Kukalova-Peck and Lawrence 2004). Historically the wing flexing mechanism (without backward folding) and the similar wing base sclerites seen in the Palaeoptera, was considered as an ancestral condition (e.g. Martynov 1925; Hennig 1969; Kukalova-Peck 1991). Furthermore, the anal brace, the intercalary veins and aquatic larvae are interpreted as plesiomorphic characters of the Ephemeroptera and Odonata (Kukalova-Peck 1991; Staniczek 2000; Bechly et al. 2001). In contrast, the suppression of imaginal molts, the absence of the axillar-furcal muscle, the basalar-sternal muscles and the missing terminal-filum observed in the Odonata and Neoptera are possible synapomorphies supporting the Metapterygota scenario (e.g. Kristensen 1991; Beutel and Gorb 2001; Grimaldi and Engel 2005; Willkommen and Hornschemeyer 2007). Alternately the direct sperm transfer shared by the Ephemeroptera and Neoptera in contrast to the indirect sperm transfer in Odonata support the Chiasmomyaria theory (Boudreaux 1979). Moreover the wing base structure of the Odonata and the remaining pterygote orders show significant differences in appearance and function, e.g. wing flapping in Odonata is promoted by the direct flight muscles whereas in Ephemeroptera and Neoptera it is promoted by indirect flight muscles (Ninomiya and Yoshizawa 2009). The difficulties in establishing homology of the wing base structure between the Odonata and other Pterygota resulted in an extreme interpretation of Matsuda (1970; 1981) and La Greca (1980). They concluded that the wing base structure in odonates is so different that it cannot be homologized with that of Ephemeroptera and Neoptera. However, the monophyly of Pterygota is now well

established through both morphology and molecular data (e.g. Kristensen 1991; Wheeler et al. 2001; Grimaldi and Engel 2005; Kjer et al. 2006; von Reumont et al. 2009). Recently Ninomiya and Yoshizawa (2009) established the homology of the wing base structures between the Odonata, Ephemeroptera and Neoptera. Based on wing base morphology, they almost unambiguously determined that there is a single origin of insect wings and flight, but were not able to contribute further on the basal diversification of Pterygota.

Establishing a sound phylogenetic hypothesis for the origin of insect wings based on wing base structure and the wing folding mechanism remains crucial and also molecular data have shown their limits to end the discussion concerning the basal diversification in Pterygota.

But why is the so-called “Palaeoptera Problem” not resolved despite the advances in molecular systematics? Whitfield and Kjer (2008) pointed out that the “ancient rapid radiation” is a major contributing factor in the inability to resolve insect relationships with molecular data. Due to short ancient internodes connecting the taxonomic groups, inadequate molecular data sets, conflicting results within or among datasets and an overall weak phylogenetic signal is observed in many pterygote phylogenetic studies (Wheeler et al. 2001; Ogden and Whiting 2003; Kjer et al. 2006; Misof et al. 2007; von Reumont et al. 2009). In addition, one major challenge is to find useful molecular markers to accurately track these short ancient internodes. For the reconstruction of an “accurate” phylogeny, molecular marker systems are required which have kept pace with speciation, but slow enough to have transferred the phylogenetic signal to the present (Regier and Shultz 1998). Unfortunately, the rationale behind the selection of certain molecular markers is not always clear, and discrepancies and incongruence between individual gene trees may result in unresolved phylogenetic trees (Wheeler et al. 2001; Kjer et al. 2006). Thus, phylogenetic analyses of single genes and even multiple marker systems have not yet conclusively resolved the basal pterygote diversification. It is therefore conceivable that resolution of these relationships may require not only large amounts of sequence data but also an assessment of data quality and quantity. Several studies have shown that analyzing a large number of genes simultaneously helps to infer unresolved issues in deep metazoan relationships (e.g. Philippe et al. 2005b; Savard et al. 2006; Roeding et al. 2007; Dunn et al. 2008). Moreover, simulations and studies based on real data have shown that trees based on concatenated alignments provide better resolution for a particular topology than consensus gene trees – known as “supertree” approaches (Rokas et al. 2003b; Gadagkar et al. 2005; Savard et al. 2006). However, there is still a controversy about phylogenetic



reconstructions derived from “supertree” versus “supermatrix” approaches (e.g. Gatesy et al. 2004; Wilkinson et al. 2007). Both methods have demonstrated strengths and weaknesses and some promising new approaches are addressing the existing problems. For example, for the supertree method the recent proposal of a maximum likelihood approach forms an important idea for future phylogenetic inferences from genomic data (Steel and Rodrigo 2008; Cotton and Wilkinson 2009). Also the implementation of new methods, e.g. BEST (Bayesian Estimation of Species Trees) (Edwards et al. 2007; Liu and Pearl 2007) to simultaneously estimate gene trees and species trees from multilocus data using a coalescent framework has been shown to be very efficient in cases of recent speciation (Edwards et al. 2007; Belfiore et al. 2008; Wiens et al. 2008). All these phylogenomic approaches have one problem in common; although the stochastic error is dramatically reduced by using a large number of data, they are not protected against systematic errors (Phillips et al. 2004; Delsuc et al. 2005). Furthermore, systematic bias can be reinforced by increasing the number of characters resulting in a highly supported but incorrect tree (Felsenstein 1978; Jeffroy et al. 2006). Long-branch attraction coupled with taxon sampling, phylogenetic reconstruction methods and base-composition bias are all factors that are known to cause systematic errors and to be potential pitfalls when attempting to recover “the true evolutionary history of species” (Zwickl and Hillis 2002; Phillips et al. 2004; Brinkmann et al. 2005; Delsuc et al. 2005; Philippe et al. 2005a).

With the aim of addressing the origin of flying insects, we generated and analyzed expressed sequence tag (EST) data from the two basal orders of winged insects – from a mayfly (Ephemeroptera, *Baetis* sp.) and a damselfly (Odonata, *Ischnura elegans*). EST data provide a comprehensive random sample of protein coding genes and an economic way to produce a large number of sequences for phylogenetic analysis of “non-model” species, for which genome sequence projects are not yet available.

Although the EST data collection is increasing due to the tremendous recent advances in sequencing technologies and as an optimal source for multi-gene approaches, ESTs from representatives of the basal winged insect orders are still scarce.

While ESTs are a promising tool to resolve deep phylogenetic questions, there are still necessary precautions to take when handling EST data sets. The complex nature of genome evolution including gene loss, duplications, expansion of gene families and functional diversification consequently requires assignment of gene orthology when using ESTs as a source for phylogenetic analyses (Hughes et al. 2006). Furthermore ESTs represent a snapshot of gene expression within a given set of tissue, developmental stages and

environmental conditions (Rudd 2003), and the overlap of genes in the taxa may be very limited (Hughes et al. 2006).

In this study, we have assigned gene orthology using the new search algorithm HaMStR (Hidden Markov Model Based Search for Orthologs using Reciprocity) (Ebersberger et al. 2009) and constructed two alignments to evaluate support for each of the three hypotheses which explain the basal relationships of the pterygotes. Our phylogenetic analyses include representatives from pterygote and apterygote (wingless) orders. The data sets differ in their number of taxa, number of genes, the proportion of missing data and consequently the overall number of characters. The data sets were subjected to different statistical and phylogenetic analyses to test the three hypotheses and to gain more insights into the origin of flying insects. The phylogenetic information contained within the different protein coding genes represented within the data sets was also assessed.

## Materials and Methods

### *cDNA library construction, EST Processing and Sequence Alignment*

Specimens were stored in RNAlater (Qiagen) at -80°C before RNA extraction. Total RNA of *Baetis* sp. was extracted four times from two larval specimens simultaneously using Qiagen RNeasy kits and pooled afterwards. Total RNA of *Ischnura elegans* was extracted from one adult specimen using Qiagen RNeasy kits. The two RNA samples were precipitated with 0.1Vol NaAC in DEPC and 2.5Vol 100% EtOH for later construction of cDNA libraries.

The Creator<sup>TM</sup> SMART<sup>TM</sup> cDNA Library Construction Kit (Clontech) and the Trimmer Kit (Evrogen) were used for the construction of the normalized cDNA libraries following manufacturers instructions. Modifications to the protocol were made concerning the cloning vector: pal32 (Evrogen) was used for directional cloning with insertion between two SfiI sites. Plasmids were transferred via electroporation to *Escherichia coli* (strain DH10B, Invitrogen). Plasmids were isolated using the method of Hecht et al. (2006) and 5'end sequenced using BigDye V3 (ABI) and 3730XL capillary sequencer systems (ABI). The program Lucy (Chou and Holmes 2001) removed vector contaminations in the raw sequences. Additionally all sequences were screened for contamination by comparing them to the UniVec database (<http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html>) with CrossMatch (<http://www.phrap.org/phredphrapconsed.html>) and SeqClean (<http://compbio.dfci.harvard.edu/tgi/software>). The latter program was also used to remove PolyA-tails. Subsequently, ESTs with less than 100 nucleotides were discarded. Repetitive

elements were soft-masked using RepeatMasker (Smit et al. 1996) and Repbase (Jurka et al. 2005).

ESTs for each species were clustered and assembled using TGICL (Pertea et al. 2003). The resulting EST contigs were quality clipped with Lucy and again sequences of less than 100 nucleotides were removed. Afterwards the quality clipped sequences were clustered a second time.

*Baetis* sp. ESTs have been deposited in the EMBL Nucleotide Sequence Database with Accession Nos FN198828-FN203024 and *Ischnura elegans* ESTs with Accession Nos FN215340-FN219556.

Individual EST contigs were compared with the NCBI non-redundant protein database using BlastX (Altschul et al. 1997). The protein sequences of the best 25 Blast hits per contig were extracted from the database and aligned to the contig separately using GeneWise (Birney et al. 2004). The description of the protein sequence resulting in the highest GeneWise alignment score was adopted as tentative annotation.

Further ESTs were downloaded from NCBI's dbEST database. A processing analog to the procedure explained above was applied except that Lucy was not used for vector screening and quality clipping and only a single clustering step was performed.

We included 25 pterygote and three apterygote specimens in our data set (Supplementary Table S1). For each taxon, identification of orthologous genes was carried out using the HaMStR approach (Ebersberger et al. 2009) (<http://www.deep-phylogeny.org/hamstr/>) with *Anopheles gambiae*, *Apis mellifera*, *Drosophila melanogaster*, *Homo sapiens* and *Aedes aegypti* as core reference taxa and a re-blast of the candidate EST contigs against *Apis mellifera* as a reference proteome. Overall our core ortholog set encompassed 3096 clusters of orthologous genes, which were used to assign EST contigs to individual genes.

Since not all of the 3096 genes were present in the EST contigs of each taxon, a concatenation of all gene alignments would have resulted in a substantial amount of missing data. We therefore used a PERL script (kindly provided by Ingo Ebersberger and available upon request; [ingo.ebersberger@univie.ac.at](mailto:ingo.ebersberger@univie.ac.at)) that automatically analyzes the amount of missing data for different combinations of taxa and genes. We have chosen two data sets representing different taxa and genes and a diverse proportion of missing data. We decided to perform all analyses with both data sets to make our results more robust. As selection criterion of the data sets we imposed that *Baetis* sp., *Ischnura elegans* and at least one apterygote taxon was present in each set. One data set (named *maxspe*) comprised 15

species and 125 genes with 18% missing data and a second (named *maxgen*) comprised eight species, 150 genes and 11% missing data.

Sequences were aligned with MAFFT (Katoh et al. 2005) using the options --maxiterate 1000 and --localpair. Afterwards we concatenated the alignments to generate one super-alignment per data set (Supplementary Table S2 for list of represented genes).

### ***Phylogenetic analyses of concatenated data***

Both alignments (*maxspe*, *maxgen*) were checked for putative randomly similar sections using ALISCORE (Misof and Misof 2009). We applied a sliding window size ( $w=6$ ) with the BLOSUM62 matrix and function -e. After the exclusion of putative randomly similar sections using PAUP\*4.0b (Swofford 2002) we determined the best fitting model of protein sequence evolution with ProtTest 1.4 (Abascal et al. 2005) which was used in subsequent phylogenetic analyses.

*Maxspe* and *maxgen* were treated equally in all following steps of phylogenetic and statistical analyses. Tests of the three alternative phylogenetic hypotheses at the base of the Pterygota were accomplished by using the approximately unbiased test (AU), Kishino-Hasegawa (KH), Shimodaira-Hasegawa (SH), weighted Kishino-Hasegawa (WKH) and weighted Shimodaira-Hasegawa (WSH) tests as implemented in CONSEL (Shimodaira and Hasegawa 2001). First, alternative tree topologies were reconstructed by using GARLI 0.96b8 (Zwickl 2006) under default parameters. Heuristic searches were conducted assuming the WAG (Whelan and Goldman 2001) model of amino acid sequence evolution and a  $\Gamma$ -model of rate heterogeneity (Gu et al. 1995), with four classes of variable sites and one class of invariable sites ( $4\Gamma+I$ ). As the monophyly of the major groups is not disputed, we put a topological constraint according to the three phylogenetic hypotheses on the tree search to identify the highest likelihood topologies that satisfied a given hypothesis. In addition we constrained the monophyly of Paraneoptera and Holometabola in the *maxspe* data set and the monophyly of Holometabola in the *maxgen* data set (e.g. Hennig 1981; Yoshizawa and Saigusa 2001; Kaestner 2003; Beutel and Pohl 2006). Second PAUP\* was used to produce a file with the site wise log-likelihoods of alternative trees. The resulting files were summarized to a single file that served as input for CONSEL to calculate the  $p$ -value for each alternative phylogenetic hypothesis.

In addition to the constrained analyses, searches in the absence of topological constraints were carried out. For this purpose maximum likelihood analyses (ML) were performed with the Pthreads-parallelized version of RAxML 7.0.4 (Stamatakis 2006) under a rapid

bootstrap analysis (-f a) and the PROTMIXWAG model. The branching support was assessed by 1,000 bootstrap replicates. Bayesian inference (BI) analyses were performed using a compiled parallel version of MrBayes v3.1.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003; Altekar et al. 2004) with two parallel runs under the WAG+4 $\Gamma$ +I model. Metropolis-coupled Markov chain Monte Carlo (MCMCMC) sampling was carried out with one cold and three heated chains starting from random starting trees and the program default prior probabilities on model parameters. The *maxspe* data were run for 3,000,000 generations (average standard deviation of split frequencies < 0.0025), and the *maxgen* data were run for 1,000,000 generations (average standard deviation of split frequencies < 0.0000). For both data sets samples of the Markov chain were taken every 100 generations giving a total sample of 30,000 trees (*maxspe*) or 10,000 trees (*maxgen*). Parameters were checked for stationarity with Tracer v1.4 (Rambaut and Drummond 2007) and the first 10,000 trees were discarded as burn-in. Bayesian posterior probabilities were obtained from the majority rule consensus of the tree sampled after the initial burn-in period.

## Results

### *ESTs and alignments from Baetis sp. and Ischnura elegans*

After trimming of vector, filtering for minimum length (<100bp) and removal of low-quality sequences, we obtained 4,197 *Baetis* sp. sequences from the initial 4,225 clones. For *Ischnura elegans* we obtained 4,217 from 4,219 randomly sequenced clones. Clustering resulted in 3,035 contigs (635 contigs contain more than one EST, 2,400 singletons) for *Baetis* sp. and 3,194 (614 contigs contain more than one EST, 2,580 singletons) for *Ischnura elegans*.

Based on the HaMStR approach (Ebersberger et al. 2009) we identified 436 orthologous sequences in *Baetis* sp. and 527 orthologous sequences in *Ischnura elegans*.

Due to the limited number of assigned orthologs in each species, the data sets differed significantly in their represented species, genes, proportion of missing data within the taxa and their overall sequence length. In the *maxspe* alignment, 15 species and 125 genes were represented with an alignment length of 31,643aa and a proportion of missing data of 18%. The *maxgen* alignment maximized the represented genes (150) but reduced the taxa number to eight, had a sequence length of 42,541aa and a proportion of missing data of 11%. See Supplementary Table S2 for overview of represented genes in both data sets.

### Phylogenetic analyses of concatenated alignments

After the exclusion of randomly similar aligned sections identified by ALIScore (Misof and Misof 2009) the data set *maxspe* comprised 27,327aa (initial 31,643aa, ~14% randomly similar) and *maxgen* comprised 37,473aa (initial 42,541, ~12% randomly similar). The final alignments have been deposited at TREEBASE (<http://www.treebase.org>, study GenBank accession number S2456). Results of the hypotheses testing using heuristic search and incorporating topology constraints are summarized in Table 1. Based on the constrained analyses, the Chiasmomyaria scenario (Odonata, Ephemeroptera+Neoptera) is significantly supported by all tests (AU, KH, SH, WKH and WSH) in the *maxspe* data sets while the *maxgen* alignment could not significantly reject the Metapterygota theory in the weighted SH test (WSH=0.062) using the 95 percent significance level.

**Table 1**

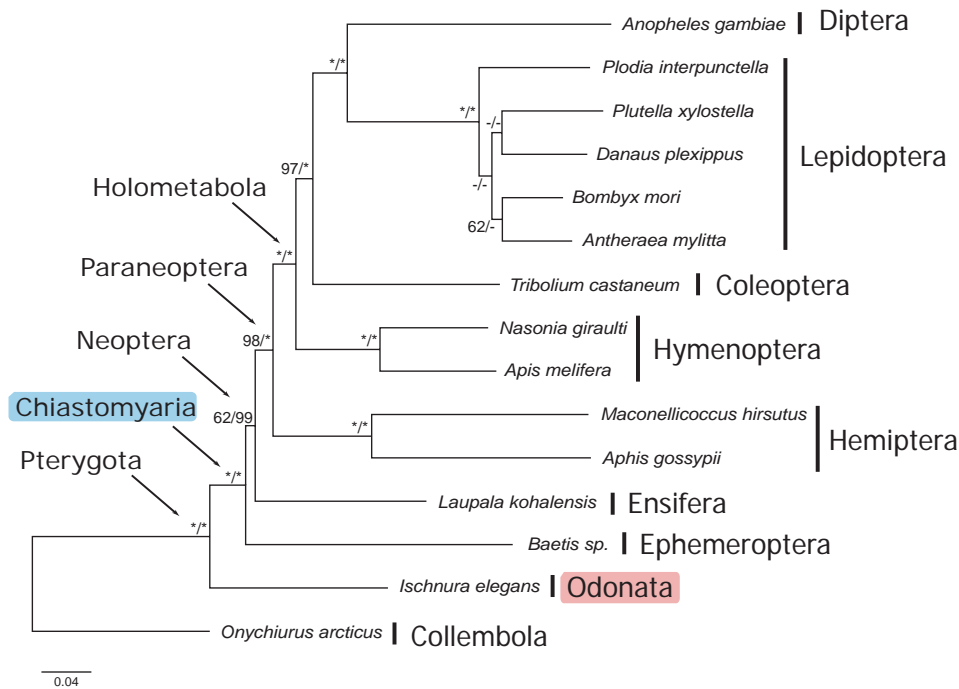
Statistical confidence (*p*-values) for alternative relationships at the base of the pterygotes. AU: approximately unbiased test, KH: Kishino-Hasegawa test, SH: Shimodaira-Hasegawa test, WKH: weighted Kishino-Hasegawa test and WSH: weighted Shimodaira-Hasegawa test.

<i>p</i> -values						
Data set	Hypothesis	AU	KH	SH	WKH	WSH
<i>maxgen</i>	Palaeoptera	2e-04**	0.001*	0.001*	0.001*	0.001*
	Metapterygota	0.032*	0.035*	0.039*	0.035*	0.062
	Chiasmomyaria	0.971	0.965	0.987	0.965	0.985
<i>maxspe</i>	Palaeoptera	7e-50***	0.002*	0.002*	0.001*	0.001*
	Metapterygota	0.029*	0.025*	0.025*	0.025*	0.048*
	Chiasmomyaria	0.971	0.975	0.982	0.975	0.979

We employed maximum likelihood (ML) and Bayesian inference (BI) analyses to construct phylogenetic trees from the *maxspe* alignment (Fig.2) and the *maxgen* alignment (Fig.3). In all trees *Ischnura elegans* (Odonata) represent – with high bootstrap support/posterior probability (*maxspe*: 100%/100%, *maxgen*: 100%/100%) – the most basal winged insect specimens, supporting the Chiasmomyaria theory. The topology generated from the *maxspe* alignment further supports the monophyly of Paraneoptera (*Aphis gossypii*, *Maconellicoccus hirsutus*) (98%/100%) and Holometabola (100%/100%), with a basal position of Hymenoptera within the Holometabola data set (Fig.2). The relationships within the Lepidoptera were not well supported in the ML (72%-32%) and the BI (88%-64%) analyses based on the *maxspe* data set.

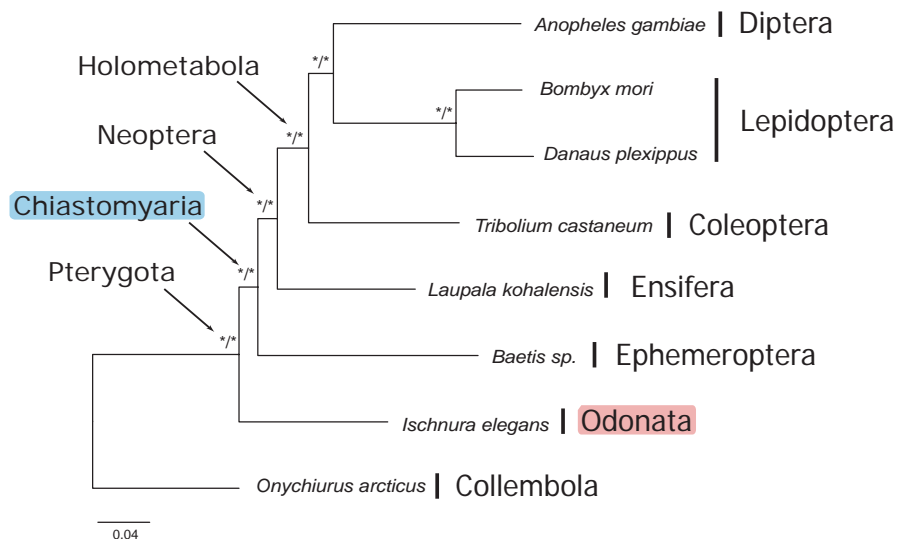
We also note that the tree based on the *maxgen* alignment is a true sub-tree of the *maxspe* tree. This indicates that the results are robust with respect to the number of species and genes. To further evaluate the quality of fit for the chosen model of evolution, we

performed the test developed by Goldman (1993). The results (see Supplementary Figures S1 and S2) support that the WAG model describes the data adequately.



**Fig. 2.** – Maximum likelihood + Bayesian inference topology of *maxspe*

Pterygote phylogenetic relationships based on 15 taxa and 125 genes data set (*maxspe*) showing a basal position of Odonata (*Ischnura elegans*), the monophyly of Paraneoptera and Holometabola. Branch lengths are from maximum likelihood trees. Bootstrap support values of maximum likelihood analysis and Bayesian posterior probabilities for each branch are indicated before and after a slash, respectively. \* indicates 100% support value, - indicates support value below 50%.



**Fig. 3.** – Maximum likelihood + Bayesian inference topology of *maxgen*

Pterygote phylogenetic relationships based on 8 taxa and 150 genes data set (*maxgen*) showing a basal position of Odonata (*Ischnura elegans*). Branch lengths are from maximum likelihood. Bootstrap support values of maximum likelihood analysis and Bayesian posterior probabilities for each branch are indicated before and after a slash, respectively. \* indicates 100% support value, - indicates support value below 50%.

### Phylogenetic analyses of single alignments

Both data sets were scanned for individual genes represented in *Baetis* sp., *Ischnura elegans* and *Onychiurus arcticus*, as well as in at least one neopterous insect. In the *maxspe* alignment we identified 39 genes and in the *maxgen* alignment 58 genes. Of these, 34 genes are present in both alignments. The function of these 63 genes was assessed through Blast against the KOG (Eukaryotic Orthologous Groups) database (<http://biotec.icb.ufmg.br/K-EST/begin.html>) and assigned to the four major KOG categories: (1) cellular processes and signaling, (2) information storage and processing, (3) metabolism and (4) poorly characterized (Table 2).

**Table 2**

Genes shared between *Baetis* sp., *Ischnura elegans* and *Onychiurus arcticus*, as well as at least one neopterous insect. These genes were assembled in the four major KOG (Eukaryotic Orthologous Groups) categories: (1) cellular processes and signaling, (2) information storage and processing, (3) metabolism and (4) poorly characterized. ID number – the numerical identifier assigned to the gene during the HaMStR process, FlyBaseID/gene name – the corresponding ID number/gene name of the *Drosophila melanogaster* genome database (<http://flybase.org/>). *maxspe*/*maxgen* – genes represented in the alignments. These genes were also selected for the extended ML analyses of individual alignments.

KOG category	ID number	FlyBaseID	gene name (FlyBase)	<i>maxspe</i>	<i>maxgen</i>
(1) cellular processes and signaling	6936	FBgn0038166	CG9588		+
	7538	FBgn0034709	CG3074	+	+
	7640	FBgn0015282	Proteasome 26S subunit 4 ATPase	+	+
	8073	FBgn0023174	Proteasome $\beta$ 2 subunit	+	+
	8075	FBgn0003150	Proteasome 29kD subunit	+	+
	8671	FBgn0033663	ERp60	+	+
	9489	FBgn0002174	lethal (2) tumorous imaginal discs		+
	8547	FBgn0010638	Sec61 $\beta$	+	
	8032	FBgn0010226	Glutathione S transferase S1	+	+
	8784	FBgn0011217	effete	+	+
	8782	FBgn0010602	lesswright		+
	9616	FBgn0037756	CG8507		+
	9827	FBgn0025637	skpA	+	+
	7864	FBgn0036928	Translocase of outer membrane 20		+
	7902	FBgn0037231	CG9779		+
	8323	FBgn0024833	AP-47		+
	9169	FBgn0021814	Vps28		+
	7720	FBgn0025700	CG5885	+	+
	9562	FBgn0028985	Serine protease inhibitor 4	+	+
	7339	FBgn0011760	cut up	+	+
	9414	FBgn0052672	Autophagy-specific gene 8a	+	+



(2) information storage and processing	9511	FBgn0014189	Helicase at 25E		+
	7970	FBgn0001197	Histone H2A variant	+	+
	7512	FBgn0037346	extra bases	+	+
(3) metabolism	6671	FBgn0029897	Ribosomal protein L17	+	+
	6790	FBgn0001942	Eukaryotic initiation factor 4a	+	+
	6906	FBgn0034967	eIF-5A	+	+
	6927	FBgn0037351	Ribosomal protein L13A	+	+
	7007	FBgn0010265	Ribosomal protein S13	+	+
	7098	FBgn0036213	Ribosomal protein L10Ab	+	+
	7316	FBgn0034743	Ribosomal protein S16	+	+
	7606	FBgn0005593	Ribosomal protein L7	+	+
	7883	FBgn0039713	Ribosomal protein S8	+	+
	7950	FBgn0034751	Ribosomal protein S24	+	+
	8013	FBgn0002590	Ribosomal protein S5a	+	+
	8023	FBgn0010409	Ribosomal protein L18A	+	+
	8456	FBgn0034138	Ribosomal protein S15	+	
	8732	FBgn0064225	Ribosomal protein L5	+	+
	8997	FBgn0039129	Ribosomal protein S19b	+	+
	9404	FBgn0031980	Ribosomal protein L36A	+	
	9821	FBgn0036825	Ribosomal protein L26	+	
	6715	FBgn0024558	Diphthamide methyltransferase		+
	9590	FBgn0028737	Elongation factor 1 $\beta$	+	+
	7383	FBgn0023211	Elongin C		+
	7771	FBgn0023212	Elongin B		+
	6637	FBgn0014028	Succinate dehydrogenase B		+
	9384	FBgn0011361	mitochondrial acyl carrier protein 1		+
	9813	FBgn0031436	CG3214		+
	9569	FBgn0028662	VhaPPA1-1		+
	7434	FBgn0039697	CG7834	+	+
	7214	FBgn0000116	Arginine kinase	+	+
	9594	FBgn0250814	CG4169		+
	6958	FBgn0036580	PDCD-5	+	+
	9095	FBgn0028833	Dak1		+
	9007	FBgn0250837	Deoxyuridine triphosphatase		+
	7631	FBgn0032192	CG5731	+	+
	8076	FBgn0033879	CG6543	+	+
(4) poorly characterized	8942	FBgn0024188	separation anxiety		+
	7015	FBgn0086254	CG6084	+	+
	7736	FBgn0035528	CG15012		+
	7742	FBgn0038739	CG4686		+
	8092	FBgn0030724	Nipsnap		+

We performed extended ML-tree analyses of the individual *maxspe* (total 39) and *maxgen* (total 58) alignments to investigate the support of the three phylogenetic hypotheses by the individual genes. The log likelihood for each topology was calculated using TREE-PUZZLE 5.2 (Schmidt et al. 2002). The topologies were considered as supported by the individual gene alignments if the  $p$ -SH  $< 0.05$  and if the  $\Delta\log L : \text{S.E.}$  ratio exceeded 0.5 (Supplementary Table S3). In addition, for each gene alignment of the *maxspe* (Supplementary Table S4a) and *maxgen* set (Supplementary Table S4b), that included a sequence of *Baetis* sp., *Ischnura elegans*, *Onychiurus arcticus* and at least one neopterous insect, a maximum likelihood tree with 100 bootstrap replicates was calculated using RAxML. Within *maxgen*, based on the  $p$ -SH value and the  $\Delta\log L : \text{S.E.}$  ratio, two genes (lethal (2) tumorous imaginal discs and Helicase at 25E) support the Metapterygota hypothesis and the gene Cysteine proteinase Cathepsin L (K-EST description) supports the Chiasmomyaria hypothesis. The majority of the genes (55) represented in the *maxgen* set did not carry sufficient phylogenetic signal to distinguish between the three alternative topologies (Supplementary Table S3). In addition, the bootstrap analyses for each gene alignment did not provide significant support ( $> 95\%$ ) for a single phylogenetic hypothesis (Supplementary Table S4a). To increase phylogenetic signal, the genes of the *maxgen* data set were concatenated according to their KOG category and subjected to ML-tree analyses using the same methods as in the individual gene analyses. Table 3 summarizes the support for the three phylogenetic hypotheses as recoded for analyses based on the functional classification using the statistical methods. The proteins involved in cellular processes and signaling (concatenated=5,285aa) gave the strongest support for the Chiasmomyaria hypothesis and rejected significantly both other topologies. The proteins contained in the metabolism category (concatenated=3,143aa) also favor the Chiasmomyaria hypothesis but did not significantly reject the Metapterygota hypothesis. Proteins classified as information storage and processing proteins (concatenated=4,697aa) favor the Metapterygota hypothesis but did not reject the Chiasmomyaria hypothesis. The poorly characterized proteins (1,179aa) identified the Metapterygota topology as the best but again did not reject the remaining hypothesis.

None of the individual *maxspe*- alignments, which were also subjected to extended ML-tree analysis using TREE-PUZZLE and RAxML, provide significant support for one of the phylogenetic hypotheses (see Supplementary Table S3 and S4a). To increase the phylogenetic signal we also concatenated the individual *maxspe*- alignments based on their KOG category assignment (cellular processes and signaling (3,511aa), information storage

and processing (4,245aa), metabolism (1,551aa) and poorly characterized (329aa)). Three of the four KOG category derived *maxspe*-alignments identified the Chiasmomyaria phylogeny as the best ML-tree, but the two alternative topologies could not be rejected by the proteins involved in information storage and processing + metabolism, while the genes involved in cellular processes and signaling significantly support the Chiasmomyaria theory. Proteins categorized as poorly characterized identified the Palaeoptera topology as the best tree but not significantly (summarized as Table 3).

**Table 3**

Maximum likelihood support for the three different phylogenetic hypotheses of the concatenated alignments based on their KOG category. The favored topology of each KOG category is indicated in bold. The support is expressed as the  $\Delta\log L$  : S.E. and the  $p$ -SH value. The  $-\log L$  value of the best tree is written in square brackets.

Data set	Hypothesis	cellular processes and signaling		information storage and processing		metabolism		poorly characterized	
		$\Delta\log L$ : S.E.	$p$ -SH	$\Delta\log L$ : S.E.	$p$ -SH	$\Delta\log L$ : S.E.	$p$ -SH	$\Delta\log L$ : S.E.	$p$ -SH
<i>maxgen</i>	Palaeoptera	6.61	<0.0000***	2.13	0.0570	0.87	0.2040	1.06	0.2040
	Metapterygota	6.64	<0.0000***	<b>[31382.03]</b>	<b>1.00</b>	2.05	0.0290	<b>[19394.10]</b>	<b>0.0290</b>
	Chiaatomyaria	<b>[47438.79]</b>	<b>1.00</b>	0.07	0.5860	<b>[24472.80]</b>	<b>1.00</b>	1.60	1.00
<i>maxspe</i>	Palaeoptera	6.79	<0.0000***	1.64	0.0910	0.74	0.2650	<b>[3228.31]</b>	<b>0.2650</b>
	Metapterygota	6.77	<0.0000***	0.37	0.4810	1.41	0.1160	1.04	0.1160
	Chiaatomyaria	<b>[46215.35]</b>	<b>1.00</b>	<b>[43602.25]</b>	<b>1.00</b>	<b>[22195.41]</b>	<b>1.00</b>	1.00	1.00

## Discussion

The question of the first winged insect order has been dominated by the analyses of morphological characters and nuclear rRNA data (18S and 28S). Recently Zhang et al. (2008) published the first mitochondrial genome of an Ephemeropteran. The analysis used the mitogenomic approach and supported the Metapterygota hypothesis. Despite numerous studies concerning the phylogenetic relationships at the base of pterygotes, the so-called “Palaeoptera problem” is still not solved and results are often conflicting. A combined analysis of nuclear rRNA (18S and 28S) and 275 morphological characters supported the Metapterygota hypothesis (Wheeler et al. 2001) as did a combined analysis of 18S+28S rRNA, the protein coding gene Histone 3 and morphology data (Ogden and Whiting 2003). This hypothesis is supported by some diagnostic morphological characters connecting Ephemeroptera with the apterygote hexapods, such as molting, muscle structure in the tracheal system and the caudal filament (Kristensen 1991). However, different analyses of nuclear rRNA data by different authors support each of the three phylogenetic hypothesis depending on the phylogenetic inference method used e.g. combined 18S and 28S supports the Metapterygota hypothesis (Wheeler et al. 2001; Ogden and Whiting 2003), the Palaeoptera hypothesis (Hovmöller et al. 2002) and the Chiasmomyaria hypothesis (Mallatt and Giribet 2006; von Reumont et al. 2009). The longest standing hypothesis and the traditional textbook scenario based on morphological characters is the Palaeoptera hypothesis. It is supported by the inability of the Ephemeroptera and Odonata to fold their wings over the abdomen (Hennig 1969; Kukalova-Peck 1991), the intercalary veins in the wings, the fusion of the galea and lacinia in the larval maxillae, and the aquatic larvae (Hennig 1981). Kjer et al. (2006) also supported this hypothesis using nine genes and 170 morphological characters. However, a strong argument for the third hypothesis – the Chiasmomyaria hypothesis – is the indirect sperm transfer mechanism linking the Odonata to the apterygote insects (Boudreaux 1979) and the direct flight muscles which are a unique character of Odonata. This phylogenetic hypothesis is further supported by several molecular studies (Kjer 2004; Yoshizawa and Johnson 2005; Misof et al. 2007).

All studies clearly illustrate that basal pterygote divergence is difficult to unveil, despite the use of various morphological characters and molecular markers. One major problem is certainly the fast evolution of the pterygotes and the enormous diversity within this group. Furthermore the preserved ancient characters in some taxa and the rate heterogeneity among orders lead to confusion among phylogeneticists. For example, Kjer et al. (2006) observed excessive substitution rate acceleration for Diptera and Diplura, while Odonata

and Mantodea seem to almost “stand still”. Finding appropriate molecular markers with phylogenetically informative sites tracking the narrow window, within which the divergence and origin of winged insects took place, is the major challenge. In this study we included two crucial new basal winged insect EST data sets (representing the Odonata and Ephemeroptera), adopted a multi-gene approach and evaluated the support of different classes of functional protein coding genes for each of the three hypotheses.

Protein coding sequences obtained by EST sequencing represent a valuable and relatively inexpensive possibility for resolving long outstanding deep phylogenetic relationships. The conserved nature of the housekeeping genes makes studies of divergences which took place millions of years ago possible. Thus, phylogenetic trees inferred from multi-gene approaches using ESTs have become a popular method to resolve long outstanding questions in deep metazoan relationships. Dunn et al. (2008) for example, improve the resolution of the animal tree of life using a concatenated alignment of 150 genes, Philippe et al. (2004) concatenated 129 orthologous proteins for eukaryotic species and Savard et al. (2006) assembled 185 genes to resolve the radiation of Holometabolous insects. The advantages of a multi-gene approach instead of a single gene or few genes are numerous. Rokas et al. (2003) pointed out that the biological process of a gene as influenced by natural selection or genetic drift may cause the history of the genes under analysis to obscure the history of the taxa. Issues such as gene duplication and lineage sorting may contribute to varying degrees of discordance between gene tree and species tree. Therefore conflicting topologies are often seen in analyses of a single or small numbers of concatenated genes. Furthermore, the use of one or a few genes is known to be insufficient for the resolution of many clades (Baptiste et al. 2002; Rokas et al. 2003a; Rokas et al. 2003b), whereas larger amounts of data and the increasing number of phylogenetic informative positions robustly resolve the topology (Philippe et al. 2004). However, is a multi-gene approach really a panacea for the accurate resolution of a species tree? A study by Gadagkar et al. (2005) indicates this may not be the case, by showing that weak phylogenetic signals can be substantially reinforced when sequences are concatenated, but in the worst case it can also enhance support for the erroneous inferences, leading to very high bootstrap support for incorrect clades. In other words, the multi-gene approach does not necessarily lead to the correct topology, because adding of new genes does not increase the accuracy of the topology in the presence of a bias. Various studies have shown that the consistency of tree-reconstruction in phylogenomic studies is sensitive to the model of sequence evolution (Phillips et al. 2004; Jeffroy et al. 2006) and to taxon sampling (Hillis

et al. 2003; Brinkmann et al. 2005), both potential sources of long-branch attraction (LBA) artifacts. Subsequently, the detection and avoidance of LBA artifacts remain the most important challenge for phylogenomic studies. One strategy to reduce the impact of systematic bias would be to apply probabilistic methods which take into account variable evolutionary rates over sites and lineages (Kolaczkowski and Thornton 2004; Brinkmann et al. 2005). Unfortunately, no current model covers the full complexity of biological history which can minimize the inconsistency of methods caused by model misspecification (Steel 2005).

In this study, we have attempted to identify the impact of systematic bias in our phylogenetic analyses by applying suitable methods of analysis to better match the data, and did not detect any severe model violations (see Supplementary Figures S1, S2, S3 and S4). Adequate taxon sampling remains the other crucial factor in phylogenomic studies to avoid long-branch attraction (LBA) artifacts. Increasing the number of ingroup taxa from 7 (*maxgen*) to 14 (*maxspe*) resulted in a congruent topology and support the Chiasmomyaria hypothesis, that is, a basal position of Odonata. However, given the existing data we are not in the position to significantly enlarge taxon sampling. At this time the Chiasmomyaria hypothesis is well supported, but we are aware that possible pitfalls (LBA, wrong model of sequence evolution, gene sampling) exist. Thus, future extended analyses are necessary to finally confirm the Chiasmomyaria hypothesis.

On the other hand, not only is the phylogenomic methodology or taxon sampling important but the genes/proteins to which it is applied are also of relevance. The evolutionary history of the genes that compose the data sets may have a direct impact on the reconstructed phylogeny (Comas et al. 2007). The phylogenetic signal of a gene is likely to be related to its evolutionary constraint and it has been suggested that a polytomy can be resolved by using genes that evolve at the optimal rate in the relevant time scale (Townsend 2007).

We therefore assessed the biological function of the represented genes and concatenated them according to their functional classification with the assumption that they harbor the same evolutionary history along the branches of the organismal phylogeny. It has been known that different evolutionary signals are a result of the different evolutionary processes that act upon the genes and that the functional role of these genes in the cell is important for the phylogenetic signal they carry (Graur and Li 2000).

The statistical tests of concatenated alignments based on their functional classification showed that proteins belonging to the cellular processes and signaling category seem to harbor the strongest phylogenetic signal for resolving deep phylogenetic relationships. Our

results are congruent with a phylogenetic study of the fungal kingdom (Kuramae et al. 2007). These authors evaluated phylogenetically informative proteins for the fungal Tree of Life and identified proteins involved in cellular processes and signaling as phylogenetically more informative than the others.

Nevertheless, the large data set based on KOG (Eukaryotic Orthologous Groups) categories (*maxgen*: cellular processes and signaling = 5,285aa, information storage and processing = 4,697aa, metabolism = 3,143aa, poorly recognized proteins = 1,179aa; *maxspe*: cellular processes and signaling = 3,511aa, information storage and processing = 4,245aa, metabolism = 1,551aa, poorly recognized = 329aa), gave in the majority of analyses no strong statistical support for either one hypotheses. There are several explanations for this observation. First of all, multiple substitutions at the same positions are expected to be frequent because the speciation event occurred millions of years ago. The saturation of the molecular markers will certainly reduce the phylogenetic signal and consequently the resolution. To investigate this, we conducted ML analyses for each protein separately using TREE-PUZZLE (WAG+4Γ+I) and RAxML (PROTMIXWAG). As expected, due to the limited number of alignment positions, the analyses from the individual alignments have shown that one gene did not harbor enough phylogenetic signal to unequivocally resolve the “Palaeoptera problem”. Although the conserved nature of housekeeping genes is beneficial to track Mesozoic divergences, the phylogenetic content of single genes is too low, while concatenation seems to compensate for this fact.

It appears that the ancient rapid radiation that took place with the transition from non-winged to winged insects represents one of the major obstacles for insect systematics. As we have shown for one of the major questions in insect phylogeny, molecular phylogenetics may overcome this hurdle by closing the gaps of genetic information from key orders, carefully applying multi-gene approaches and assessing the data quality.

### Supplementary Material

Supplementary Tables S1-S4 and Figures S1-S4 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

### Acknowledgements

This work was supported by a German Research Foundation (Deutsche Forschungsgemeinschaft [DFG]) special priority program “Deep Metazoan Phylogeny” SP1174 grant given to H.H. (DFG HA 1947/5), AvH (DFG HA 1628/9) and the



Boehringer Ingelheim Fonds (B.I.F.) given to S.Si. which supported the Workshop on Molecular Evolution 2008, Marine Biological Laboratory, Woods Hole, MA, USA. AvH and S.St. would also like to thank the WWTF for generous funding.

S.Si. would like to express her gratitude to Michael P. Cummings, Steven Thompson and Akito Y. Kawahara for helpful suggestions concerning data analyses during the MBL Workshop in Woods Hole. We thank Sara Khadjeh for providing specimens and RNA of *Ischnura elegans* and Michael Kube and Richard Reinhardt (MPI for Molecular Genetics, Berlin, Germany) for the construction and sequencing of cDNA libraries. Special thanks go also to Ingo Ebersberger (CIBIV, Vienna) for providing a pre-release of the HaMStR tool and the PERL script to calculate the amount of missing data and Danielle de Jong for linguistic help, and Bernd Schierwater for his helpful comments on an earlier version of the manuscript. We also thank Associate Editor Barbara Holland and two anonymous reviewers for providing constructive comments which greatly improved this manuscript.

### Literature Cited

- Abascal, F., R. Zardoya, and D. Posada. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21:2104-2105.
- Altekar, G., S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20:407-415.
- Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.
- Baptiste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc Natl Acad Sci U S A* 99:1414-1419.
- Bechly, G., C. Brauckmann, W. Zessin, and E. Gröning. 2001. New results concerning the morphology of the most ancient dragonflies (Insecta: Odonatoptera) from the Namurian of Hagen-Vorhalle (Germany). *J. Zool. Syst. Evol.* 39:209-226.
- Belfiore, N. M., L. Liu, and C. Moritz. 2008. Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). *Syst Biol* 57:294-310.
- Beutel, R. G., and S. Gorb. 2001. Ultrastructure of attachment specializations of hexapods (Arthropoda): evolutionary patterns inferred from a revised ordinal phylogeny. *Journal of Zoological Systematics and Evolutionary Research* 39:177-207.
- Beutel, R. G., and H. Pohl. 2006. Endopterygote systematics - where do we stand and what is the goal (Hexapoda, Arthropoda)? *Systematic Entomology* 31:202-219.
- Birney, E., M. Clamp, and R. Durbin. 2004. GeneWise and Genomewise. *Genome Res* 14:988-995.
- Boudreaux, H. B. 1979. *Arthropod phylogeny with special reference to insects*. Wiley, New York.
- Brinkmann, H., M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol* 54:743-757.

- Chou, H. H., and M. H. Holmes. 2001. DNA sequence quality trimming and vector removal. *Bioinformatics* 17:1093-1104.
- Comas, I., A. Moya, and F. Gonzalez-Candelas. 2007. Phylogenetic signal and functional categories in Proteobacteria genomes. *BMC Evol Biol* 7 Suppl 1:S7.
- Cotton, J. A., and M. Wilkinson. 2009. Supertrees join the mainstream of phylogenetics. *Trends Ecol Evol* 24:1-3.
- Delsuc, F., H. Brinkmann, and H. Philippe. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet* 6:361-375.
- Dunn, C. W., A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sorensen, S. H. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. M. Kristensen, W. C. Wheeler, M. Q. Martindale, and G. Giribet. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745-749.
- Ebersberger, I., S. Strauss, and A. von Haeseler. 2009. HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol* 9:157.
- Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. *Proc Natl Acad Sci U S A* 104:5936-5941.
- Engel, M. S., and D. A. Grimaldi. 2004. New light shed on the oldest insect. *Nature* 427:627-630.
- Felsenstein, J. 1978. Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. *Systematic Zoology* 27:401-410.
- Gadagkar, S. R., M. S. Rosenberg, and S. Kumar. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J Exp Zool B Mol Dev Evol* 304:64-74.
- Gatesy, J., R. H. Baker, and C. Hayashi. 2004. Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of Crocodylia. *Syst Biol* 53:342-355.
- Gaunt, M. W., and M. A. Miles. 2002. An insect molecular clock dates the origin of the insects and accords with palaeontological and biogeographic landmarks. *Mol Biol Evol* 19:748-761.
- Goldman, N. 1993. Statistical tests of models of DNA substitution. *J Mol Evol* 36:182-198.
- Graur, D., and W. H. Li. 2000. *Fundamentals of molecular evolution*. Sinauer Associates, Sunderland, Massachusetts.
- Grimaldi, D. A., and M. S. Engel. 2005. *Evolution of the Insects*. Cambridge University Press, New York.
- Gu, X., Y. X. Fu, and W. H. Li. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol Biol Evol* 12:546-557.
- Hecht, J., H. Kuhl, S. A. Haas, S. Bauer, A. J. Poustka, J. Lienau, H. Schell, A. C. Stiege, V. Seitz, R. Reinhardt, G. N. Duda, S. Mundlos, and P. N. Robinson. 2006. Gene identification and analysis of transcripts differentially regulated in fracture healing by EST sequencing in the domestic sheep. *BMC Genomics* 7:172.
- Hennig, W. 1969. *Die Stammesgeschichte der Insekten*. Senckenbergische Naturforschende Gesellschaft, Frankfurt am Main.
- Hennig, W. 1981. *Insect Phylogeny*. John Wiley & Sons, Bath, UK.
- Hillis, D. M., D. D. Pollock, J. A. McGuire, and D. J. Zwickl. 2003. Is sparse taxon sampling a problem for phylogenetic inference? *Syst Biol* 52:124-126.
- Hovmöller, R., T. Pape, and M. Källersjö. 2002. The Palaeoptera Problem: Basal Pterygote Phylogeny Inferred from 18S and 28S rDNA Sequences. *Cladistics* 18:313-323.
- Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754-755.

- Hughes, J., S. J. Longhorn, A. Papadopoulou, K. Theodorides, A. de Riva, M. Mejia-Chang, P. G. Foster, and A. P. Vogler. 2006. Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles). *Mol Biol Evol* 23:268-278.
- Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? *Trends Genet* 22:225-231.
- Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462-467.
- Kaestner, A. 2003. *Lehrbuch der Speziellen Zoologie*. Spektrum Akad. Verl., Heidelberg.
- Katoh, K., K. Kuma, H. Toh, and T. Miyata. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511-518.
- Kjer, K. M. 2004. Aligned 18S and insect phylogeny. *Syst Biol* 53:506-514.
- Kjer, K. M., F. L. Carle, J. Litman, and J. Ware. 2006. A molecular phylogeny of Hexapoda. *Arthropod Systematics & Phylogeny* 64:35-44.
- Kolaczowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431:980-984.
- Kristensen, N. P. 1991. Phylogeny of extant hexapods *in* I. D. Naumann, Lawrence, J.F., Nielsen, E.S., Spradberry, J.P., Taylor, R.W., Whitten, M.J., Littlejohn, M.J., ed. *The Insects of Australia: A Textbook for Students and Research Workers*. CSIRO, Melbourne Univ. Press, Melbourne.
- Kukalova-Peck, J. 1991. Fossil history and the evolution of hexapod structures. Pp. 141-179 *in* I. D. Naumann, Lawrence, J.F., Nielsen, E.S., Spradberry, J.P., Taylor, R.W., Whitten, M.J., Littlejohn, M.J., ed. *The Insects of Australia: A Textbook for Students and Researcher Workers*. CSIRO Melbourne University Press, Melbourne.
- Kukalova-Peck, J., and J. F. Lawrence. 2004. Relationships among coleopteran suborders and major endoneopteran lineages: Evidence from hind wing characters. *Eur J Entomol* 101:95-144.
- Kuramae, E. E., V. Robert, C. Echavarri-Erasun, and T. Boekhout. 2007. Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom. *BMC Evol Biol* 7:134.
- La Greca, M. R. 1980. Origin and evolution of wings and flight in insects. *Bollettino Zoologici* 47(Suppl):65-82.
- Liu, L., and D. K. Pearl. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst Biol* 56:504-514.
- Mallatt, J., and G. Giribet. 2006. Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. *Mol Phylogenet Evol* 40:772-794.
- Martynov, A. V. 1925. Über zwei Grundtypen der Flügel bei den Insekten und ihre Evolution. *Zeitschrift für Morphologie und Ökologie der Tiere* 4:465-501.
- Matsuda, R. 1970. Morphology and evolution of the insect thorax. *Memoirs of the Canadian Entomological Society* 76:1-483.
- Matsuda, R. 1981. The origin of insect wings (Arthropoda: Insecta). *International Journal of Insect Morphology and Embryology* 10:387-398.
- Misof, B., and K. Misof. 2009. A Monte Carlo Approach Successfully Identifies Randomness in Multiple Sequence Alignments : A More Objective Means of Data Exclusion. *Systematic Biology* 58:21-34.
- Misof, B., O. Niehuis, I. Bischoff, A. Rickert, D. Erpenbeck, and A. Staniczek. 2007. Towards an 18S phylogeny of hexapods: accounting for group-specific character

- covariance in optimized mixed nucleotide/doublet models. *Zoology (Jena)* 110:409-429.
- Ninomiya, T., and K. Yoshizawa. 2009. A revised interpretation of the wing base structure in Odonata. *Systematic Entomology* 34:334-345.
- Ogden, T. H., and M. F. Whiting. 2003. The problem with "the Paleoptera Problem:" sense and sensitivity. *Cladistics* 19:432-442.
- Perteau, G., X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, J. Tsai, and J. Quackenbush. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* 19:651-652.
- Philippe, H., F. Delsuc, H. Brinkmann, and N. Lartillot. 2005a. Phylogenomics. *Annual Review of Ecology, Evolution and Systematics* 36:541-562.
- Philippe, H., N. Lartillot, and H. Brinkmann. 2005b. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol* 22:1246-1253.
- Philippe, H., E. A. Snell, E. Baptiste, P. Lopez, P. W. Holland, and D. Casane. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol* 21:1740-1752.
- Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol* 21:1455-1458.
- Rambaut, A., and A. J. Drummond. 2007. Tracer v1.4. Available from <http://beast.bio.ed.ac.uk/Tracer>.
- Regier, J. C., and J. W. Shultz. 1998. Molecular phylogeny of arthropods and the significance of the Cambrian explosion for molecular systematics. *Am. Zool* 38:918-928.
- Roeding, F., S. Hagner-Holler, H. Ruhberg, I. Ebersberger, A. von Haeseler, M. Kube, R. Reinhardt, and T. Burmester. 2007. EST sequencing of Onychophora and phylogenomic analysis of Metazoa. *Mol Phylogenet Evol* 45:942-951.
- Rokas, A., N. King, J. Finnerty, and S. B. Carroll. 2003a. Conflicting phylogenetic signals at the base of the metazoan tree. *Evol Dev* 5:346-359.
- Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003b. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798-804.
- Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572-1574.
- Rudd, S. 2003. Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci* 8:321-329.
- Savard, J., D. Tautz, S. Richards, G. M. Weinstock, R. A. Gibbs, J. H. Werren, H. Tettelin, and M. J. Lercher. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Res* 16:1334-1338.
- Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502-504.
- Shimodaira, H., and M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246-1247.
- Smit, A. F., R. Hubley, and P. Green. 1996. RepeatMasker. unpublished.
- Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688-2690.
- Staniczek, A. H. 2000. The mandible of silverfish (Insecta: Zygentoma) and mayflies (Ephemeroptera): its morphology and phylogenetic significance. *Zoologischer Anzeiger* 239:147-178.

- Steel, M. 2005. Should phylogenetic models be trying to 'fit an elephant? *Trends in Genetics* 21:307-309.
- Steel, M., and A. Rodrigo. 2008. Maximum likelihood supertrees. *Syst Biol* 57:243-250.
- Swofford, D. L. 2002. *PAUP\* Phylogenetic Analysis Using Parsimony (\*and Other Methods)*. Sinauer Associates, Sunderland, MA.
- Townsend, J. P. 2007. Profiling phylogenetic informativeness. *Systematic Biology* 56:222-231.
- von Reumont, B. M., K. A. Meusemann, N. U. Szucsich, E. Dell'Ampio, V. Gowri-Shankar, D. Bartel, S. Simon, H. O. Letsch, R. R. Stocsits, Y. Luan, J. W. Wägele, G. Pass, H. Hadrys, and B. Misof. 2009. Can comprehensive background knowledge be incorporated into substitutions models to improve phylogenetic analyses? A case study on major arthropod relationships. *BMC Evol Biol* forthcoming.
- Wheeler, W. C., M. F. Whiting, Q. D. Wheeler, and J. M. Carpenter. 2001. The Phylogeny of the Extant Hexapod Orders. *Cladistics* 17:113-169.
- Whelan, S., and N. Goldman. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18:691-699.
- Whitfield, J. B., and K. M. Kjer. 2008. Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annu Rev Entomol* 53:449-472.
- Wiens, J. J., C. A. Kuczynski, S. A. Smith, D. G. Mulcahy, J. W. Sites, Jr., T. M. Townsend, and T. W. Reeder. 2008. Branch lengths, support, and congruence: testing the phylogenomic approach with 20 nuclear loci in snakes. *Syst Biol* 57:420-431.
- Wilkinson, M., J. A. Cotton, F. J. Lapointe, and D. Pisani. 2007. Properties of supertree methods in the consensus setting. *Syst Biol* 56:330-337.
- Willkommen, J., and T. Hornschemeyer. 2007. The homology of wing base sclerites and flight muscles in Ephemeroptera and Neoptera and the morphology of the pterothorax of *Habroleptoides confusa* (Insecta: Ephemeroptera: Leptophlebiidae). *Arthropod Struct Dev* 36:253-269.
- Yoshizawa, K., and K. P. Johnson. 2005. Aligned 18S for Zoraptera (Insecta): phylogenetic position and molecular evolution. *Mol Phylogenet Evol* 37:572-580.
- Yoshizawa, K., and T. Saigusa. 2001. Phylogenetic analysis of paraneopteran orders (Insecta: Neoptera) based on forewing base structure, with comments on monophyly of Auchenorrhyncha (Hemiptera). *Systematic Entomology* 26:1-13.
- Zhang, J., C. Zhou, Y. Gai, D. Song, and K. Zhou. 2008. The complete mitochondrial genome of *Parafronurus youi* (Insecta: Ephemeroptera) and phylogenetic position of the Ephemeroptera. *Gene* 424:18-24.
- Zwickl, D. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. The University of Texas at Austin.
- Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Syst Biol* 51:588-598.

## A phylogenomic approach to resolve the arthropod tree of life

**Karen Meusemann,<sup>\*,1</sup> Björn M. von Reumont,<sup>\*,1</sup> Sabrina Simon,<sup>2</sup> Falko Roeding,<sup>3</sup>  
Sascha Strauss,<sup>4</sup> Patrick Kück,<sup>1</sup> Ingo Ebersberger,<sup>4</sup> Manfred Walz,<sup>5</sup> Günther Pass,<sup>5</sup>  
Sebastian Breuers,<sup>6</sup> Viktor Achter,<sup>6</sup> Arndt von Haeseler,<sup>4</sup> Thorsten Burmester,<sup>3</sup> Heike  
Hadrys,<sup>2,7</sup> J. Wolfgang Wägele,<sup>1</sup> and Bernhard Misof<sup>\*,3,†</sup>**

<sup>1</sup>*Zoologisches Forschungsmuseum Alexander Koenig, Molecular Biology Unit, Adenauerallee 160, 53113 Bonn, Germany;*

<sup>2</sup>*Stiftung Tierärztliche Hochschule Hannover, ITZ, Ecology & Evolution, Buenteweg 17d, 30559 Hannover, Germany;*

<sup>3</sup>*Biozentrum Grindel & Zoologisches Museum, Martin-Luther-King Platz 3, 20146 Hamburg, Germany;*

<sup>4</sup>*Center for Integrative Bioinformatics Vienna, Max F Perutz Laboratories, University of Vienna, Medical University of Vienna, Dr.-Bohrgasse 3-9, 1030 Wien, Austria;*

<sup>5</sup>*Faculty of Life Sciences, University of Vienna, Althanstr. 14, 1090 Wien, Austria;*

<sup>6</sup>*Regional Computing Center of Cologne (RRZK), Robert-Koch-Str. 10, 50931 Köln, Germany; and*

<sup>7</sup>*Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06520, USA*

\* These authors contributed equally to this work.

† Corresponding author

This is the author's version of a work originally published by the Oxford University Press in: *Molecular Biology and Evolution* (Epub 2010 June 9), available under doi:10.1093/molbev/msq130

**Abstract**

Arthropods were the first animals to conquer land and air. They encompass more than three quarters of all described living species. This extraordinary evolutionary success is based on an astoundingly wide array of highly adaptive body organizations. A lack of robustly resolved phylogenetic relationships, however, currently impedes the reliable reconstruction of the underlying evolutionary processes. Here, we show that phylogenomic data can substantially advance our understanding of arthropod evolution and resolve several conflicts among existing hypotheses. We assembled a data set of 233 taxa and 775 genes from which an optimally informative data set of 117 taxa and 129 genes was finally selected using new heuristics and compared to the unreduced data set. We included novel EST data for eleven species and all published phylogenomic data augmented by recently published EST data on taxonomically important arthropod taxa. This thorough sampling reduces the chance of obtaining spurious results due to stochastic effects of undersampling taxa and genes. Orthology prediction of genes, alignment masking tools, and selection of most informative genes due to a balanced taxa-gene ratio using new heuristics were established. Our optimized data set robustly resolves major arthropod relationships. We received strong support for a sister group relationship of onychophorans and euarthropods, and strong support for a close association of tardigrades and cycloneuralia. Within pancrustaceans, our analyses yielded paraphyletic crustaceans and monophyletic hexapods, and robustly resolved monophyletic endopterygote insects. However, our analyses also showed for few deep splits that were recently thought to be resolved, for example the position of myriapods, a remarkable sensitivity to methods of analyses.

*Key words:* arthropod phylogeny, phylogenomics, Expressed Sequence Tags (ESTs), supermatrix, matrix saturation, relative informativeness.

## Introduction

Extensive sequence data from genome and expressed sequence tag (EST) projects were recently used to infer a deep metazoan phylogeny (Bourlat et al., 2006; Roeding et al., 2007; Delsuc et al., 2008; Dunn et al., 2008; Hejnol et al., 2009; Philippe et al., 2009). These phylogenomic studies consistently place arthropods within the superphylum Ecdysozoa. These studies are, however, sparse in their sampling of arthropods. Large groups like pancrustaceans are represented by only a few taxa, and important taxa from chelicerates, myriapods, crustaceans or hexapods are completely missing. EST studies presenting a broader arthropod taxon sampling focus on pancrustacean and hexapod relationships (Timmermans et al., 2008) or on relationships within pterygote insects (Simon et al., 2009). Other studies are essentially restricted to multi-gene analyses comprising larger arthropod data sets. Regier et al. (2008) analyzed 62 arthropod taxa covered by mainly three genes, but only for a small subset of 13 taxa were all 68 gene regions present. This multi-gene matrix, however, had 71% missing data. A large proportion of missing data within a supermatrix might cause problems for phylogenetic inference (Sanderson, 2007; Wiens and Moen, 2008). The most recent study (Regier et al., 2010) relies on selected 62 nuclear protein coding genes for 75 arthropod taxa. Important taxa assumed to be positioned at basal splits, like proturans (Hexapoda), are still missing and their data set at an amino acid level is relatively small (ca. 13,000 amino acids). Much attention was drawn to large arthropod data sets inferred from rRNA genes (Mallatt and Giribet, 2006; von Reumont et al., 2009). Drawbacks of the rRNA based studies include a lack of robust signal or conflicts in the data (see von Reumont et al., 2009). Despite this recent progress, these studies fail to completely resolve the arthropod tree of life, leaving many important questions open.

To alleviate the limitations of previous studies, we compiled a more comprehensive set of 233 taxa (214 euarthropod taxa plus 3 onychophorans, 2 tardigrades and 14 outgroup taxa) and 775 putative orthologous genes which cover 350,356 amino acid positions. We contribute data of eleven new EST projects from velvet worms, millipedes, sea spiders, barnacles, copepods, branchiopods, proturans, diplurans, springtails and bristletails. Recently published data on dragon-and mayflies (Simon et al., 2009) were also added. These thirteen projects fill critical gaps in the published data (table 1). Previous phylogenomic analyses have shown that beside massive accumulation of data, several additional elements must be part of the analysis pipeline: careful selection of orthologs,



consideration of data quality, reduction of data gappiness and model fitting must be part of the analysis pipeline (Roeding et al., 2007; Dunn et al., 2008; Hartmann and Vision, 2008; Philippe et al., 2009). Consequently, we used recently developed tools for ortholog gene prediction (Ebersberger et al., 2009, see Supplementary figure 1) and alignment masking (Misof and Misof, 2009) which facilitate a completely reproducible data analysis. Moreover, we applied new heuristics of selecting an optimal data set from a supermatrix to increase the number of taxa with potentially informative genes (Supplementary figure 2); this contrasts with other recent studies (Dunn et al., 2008; Regier et al., 2010) that rely on presence-absence matrices. The logic behind our approach is to reduce effects of poorly represented taxa and of uninformative genes by identifying and filtering these prior to tree reconstruction (see Methods and Supplementary figures 3-5). This pre-processing improves the signal-to-noise ratio in the data and considerably helped to reduce the effort spent in tree reconstructions. Retention of taxa and genes in the supermatrix was based on their contribution to the overall informativeness and the saturation of the data matrix prior to tree reconstructions, thus allowing a better exploration of tree space.

## Methods

### Molecular techniques

For thirteen arthropod species, cDNA libraries were constructed. Total RNA was prepared with standard kits from tissue or complete specimens preserved in RNA later or liquid nitrogen and stored at  $-80^{\circ}\text{C}$ , or total RNA was directly prepared from living specimens using Urea-phenol following Holmes and Bonner (1973) (table 1). For crustaceans and apterygote hexapods, RNA preparation was conducted by the Max Planck Institute for Molecular Genetics (MPIMG), Berlin, Germany. The cDNA libraries were constructed using CloneMiner (Invitrogen) or Creator SMART (Clontech, Heidelberg, Germany) at the MPIMG; cDNA libraries for pterygote insects were normalized (Simon et al., 2009). From cDNA libraries, ESTs were generated by sequencing clones from the 5' end on the automated capillary sequencer system ABI 3730XL (Applied Biosystems, Darmstadt, Germany) using BIGDYE chemistry (Applied Biosystems). Between 3,930 and 10,476 sequences were processed from cDNA libraries (table 1). All single EST sequences were deposited in EMBL (<http://www.ebi.ac.uk/embl/>) after being quality checked and assembled into unique transcripts (contigs), whereby two projects on pterygote insects originally sequenced for this arthropod study have recently been published (Simon et al., 2009).

**Table 1**

Species for novel EST data in the present study. Accession no.: Accession numbers; Proc. EST sequences: number of ESTs after processing

Species	Group	Accession no.	RNA extraction	cDNA library construction	No. of EST raw data	Proc. EST sequences	No. of EST contigs
<i>Peripatopsis sedgwicki</i>	ON	FN232766-FN243241	Urea-phenol	CloneMiner	10,611	10,476	3,452
<i>Endeis spinosa</i>	CH, Pycnogonida	FN211278-FN215339	Urea-phenol	CloneMiner	4,063	4,062	2,672
<i>Limulus polyphemus</i>	CH, Xiphosura	FN224411-FN232765	Urea-phenol	Creator SMART	8,435	8,355	4,050
<i>Archispirostreptus gigas</i>	MY, Diplopoda	FN194820-FN198827	Urea-phenol	Creator SMART	4,032	4,008	2,299
<i>Pollicipes pollicipes</i>	CR, Cirripedia	FN243242-FN247432	Absolutely RNA (Stratogene)	CloneMiner	4,224	4,191	1,721
<i>Tigriopus californicus</i>	CR, Copepoda	FN247433-FN252183	Trizol (Invitrogen)	Creator SMART	5,024	5,006	2,598
<i>Triops cancriformis</i>	CR, Branchiopoda	FM868344-FM872274	Trizol (Invitrogen)	Creator SMART	3,981	3,930	2,542
<i>Acerentomon franzi</i>	HE, Protura	FN186135-FN190445	Absolutely RNA (Stratogene)	CloneMiner	4,600	4,565	1,995
<i>Campodea cf. Fragilis</i>	HE, Diplura	FN203025-FN211277	Absolutely RNA (Stratogene)	CloneMiner	8,375	8,253	6,407
<i>Anurida maritima</i>	HE, Collembola	FN190447-FN194819	Trizol (Invitrogen)	Creator SMART	4,391	4,373	3,504
<i>Lepismachilis y-signata</i>	HE, Archaeognatha	FN219557-FN224410	Absolutely RNA (Stratogene)	CloneMiner	4,895	4,854	2,288
<i>Ischnura elegans</i> <sup>a</sup>	HE, Odonata	FN215340-FN219556	RNAeasy (Qiagen)	Creator SMART	4,219	4,217	3,194
<i>Baetis</i> sp. <sup>a</sup>	HE, Ephemeroptera	FN198828-FN203024	RNAeasy (Qiagen)	Creator SMART	4,225	4,197	3,035

ON: Onychophora; CH: Chelicerata; MY: Myriapoda; CR: Crustacea; HE: Hexapoda. <sup>a</sup> (Simon et al., 2009)

### Sequence processing and orthology assignment

We pre-processed new EST data (table 1) with LUCY (Chou and Holmes, 2001). EST data available for 190 additional euarthropods (myriapods, chelicerates, pancrustaceans) plus two onychophorans, two tardigrades and selected species of nematodes, annelids and molluscs (in total 216 species) were extracted from public databases, dbEST (NCBI), the Gene Index Project or the NCBI Trace Archive (Supplementary table 1). We screened all EST sequences for contamination and low-quality ends of sequences. Subsequently, overlapping ESTs from the same taxon were assembled into contigs using the TGICL package (Pertea et al., 2003). For the orthology prediction with HaMStR (Ebersberger et al., 2009), all contigs were translated into amino acid sequences in all reading frames. In total, 244 species were 'hamstred', of which 28 species were 'proteome' species. Thirteen species were used as primer taxa (Supplementary figure 1 and Supplementary table 1). Sequences of vertebrate species were additionally used to train hidden Markov Models (Ebersberger et al., 2009) but excluded in further phylogenetic analyses for computational reasons. Eight *Drosophila* 'proteome' species were also excluded for computational

reasons. The HaMStR search identified 775 putative orthologous genes for our original data set (233 species).

#### Alignments and alignment masking

Inferred amino acid sequences of all 775 putative orthologous genes were aligned (Supplementary figure 2) with MAFFT *L-INSI* (Katoh and Toh, 2008). The data set comprised 222 euarthropods, 3 onychophorans, 2 tardigrades, 3 vertebrates, 8 nematodes, 3 annelids and 3 molluscs. Excluding randomly similar, aligned sections can make phylogenetic analyses more reliable prior to tree reconstruction (Castresana, 2000; Misof and Misof, 2009; Kück et al., 2010). We therefore identified randomly similar sections for all gene alignments separately for each of the 775 genes with ALISCORE on the amino acid level (Misof and Misof, 2009; Kück et al., 2010) using default settings and maximal number of pairwise comparisons. In total, 57.62% of originally 826,633 amino acid positions were excluded to increase the signal-to-noise ratio. For each gene, only sequences comprising more than one half of the sequence information were included in the ALISCORE analyses. We masked each alignment with ALICUT (<http://www.utilities.zfmk.de>) by excluding all randomly similar alignment positions. All masked alignments were concatenated to a masked superalignment comprising 233 taxa and 350,356 amino acid positions.

#### Selecting an optimal subset (SOS) using new reduction heuristics

With the software MARE (MAtrix REduction) (<http://mare.zfmk.de>) the relative informativeness of each single gene within a superalignment was calculated based on weighted geometry quartet mapping (Nieselt-Struwe and von Haeseler, 2001), extended to amino acid data. Each gene received a value of informativeness between 0.0 and 1.0, reflecting the relative number of resolved quartet trees (Supplementary figure 3). Relative information content of each gene was calculated as the average value over all taxa including missing taxa. A data availability matrix indicating present (1) and absent (0) genes was then transformed into a matrix of potential information content of each taxon and gene by multiplying availability (0|1) with scores of informativeness. The total average information content (relative informativeness) of a supermatrix was calculated as the sum over all genes (see Supplementary figure 4). To select an optimal subset of taxa and genes with high total average information content, we used a simple hill climbing procedure. Reduction starts with dropping either taxon (row) or gene (column) with the lowest average information content, generating a new matrix. In case of ties, genes are excluded.

Consequently, taxa or genes with lowest average information content will be discarded from the matrix, yielding a selected optimal subset (SOS) with increased relative information content (Supplementary figure 5). We defined the copepod *Tigriopus* and the chilopod *Scutigera* as taxon-constraints; thus, they were not dropped from the submatrix. Copepods are discussed as a sister group to hexapods (Mallatt and Giribet, 2006; von Reumont et al., 2009), and *Scutigera* was the only representative of chilopods (Myriapoda). Therefore, we constrained matrix reduction to retain both species as key taxa. In order to reach an optimum of matrix reduction, we defined an optimality function  $f(\mathbf{P})$ , which takes into account that size reduction of an original matrix  $B$  and low average informativeness of a reduced matrix  $B'$  are penalized

$$f(\mathbf{P}) = 1 - |(\lambda - \mathbf{P}^{\alpha \times (1-\mathbf{P})})| \text{ if } \mathbf{P} < 1 \quad (1)$$

with  $\alpha$  as a scaling factor (default set to  $\alpha = 3$ ),  $\lambda$  as the size ratio between reduced  $B'$  and original matrix  $B$  (matrix size defined as #taxa x #genes).  $\mathbf{P}$  is maximized, if  $\mathbf{P} = 1$ , reduction stops. The optimality function favors reduction of matrices to high average information content. The connectivity between taxa was set to a minimum number of two overlapping genes and taxa. This means that two sets of taxa must share at least two taxa with both genes. Finally, the original superalignment was rewritten based on the selected optimal subset (SOS). Details of the new reduction algorithm will be published elsewhere (Misof et al., unpubl.).

### Phylogenetic analyses

We conducted ML analyses using RAxML Pthreads 7.0.0 (Stamatakis, 2006b; Ott et al., 2007) for a) the original data set (original supermatrix) comprising 233 taxa, 775 genes and 350,356 amino acid positions and b) the selected optimal subset (SOS) comprising 117 taxa and 129 genes with an alignment length of 37,476 amino acid positions. The final alignments have been deposited at Treebase (<http://purl.org/phylo/treebase/phyloids/study/TB2:S10507>).

We applied ML tree search and rapid bootstrapping within one step (-f a, 1,000 bootstrap replicates) on the SOS. For the original concatenated supermatrix, we conducted ten single ML tree searches and separate bootstrapping (100 replicates). We chose the ML tree with the best likelihood value to plot bootstrap values (Supplementary figure 6). All ML analyses were calculated with the PROTMIX (Stamatakis, 2006a) substitution model and the WAG matrix (Whelan and Goldman, 2001).

Bayesian analyses for the selected optimal subset were inferred using PhyloBayes version 2.3c (Lartillot et al., 2008) running the CAT mixture model (Lartillot and Philippe, 2004). We ran 25 MCMC chains for 20,000 cycles each, sampling every cycle. All parameter values were checked for convergence to define the burn-in (5,000 cycles). To infer a majority rule consensus (mrc) tree, we checked the discrepancy observed across all bipartitions (*maxdiff* value) of all chains by pairwise comparison and comparing 'triple' chain-combinations with the *bpcomp* tool. Harmonic means of the likelihood values of each chain (burnin excluded) were calculated. To infer the Bayesian mrc tree, we included three chains showing the lowest *maxdiff* value (0.186) while featuring the best likelihood values (harmonic means) of all 'triple-chain combinations' (table 2). All trees were rooted with Mollusca.

To identify 'unstable' taxa, we calculated leaf stability indices (Thorley and Wilkinson, 1999) from the collected bootstrap trees of the ML analysis using Phyutility (Smith and Dunn, 2008). We defined a threshold of < 95% as 'unstable'. All analyses ran for several months on Linux Clusters, HP ProLiant DL380 G5 blades (Dual quad core Intel Xeon E5345, 2.33 GHz, 2x 4MB L2-cache, 1333 MHz Bus, 32 GB RAM), of the ZFMK (molecular unit) and the RRZK (Regional Computing Center of Cologne) utilizing HPC resources (HP ProLiant, Dual quad core Intel Xeon E5345, 2.33 GHz, 2x 4MB L2-cache, 1333 MHz Bus, 32 GB RAM). RRZK resources were provided by the SuGI (Sustainable Grid Infrastructure) project (Project leader: V. Achter, University of Cologne funded by the BMBF).

**Table 2**

Log Likelihood values (harmonic means) and chain combinations of all PhyloBayes runs for the selected optimal subset (SOS). log Likelihood (harmonic means) of all log likelihood values, 20,000 cycles per chain, burn-in (5,000 cycles) excluded; chain combination consisting of three chains each per combination (triple) for which the *maxdiff* value < 0.3; *maxdiff*: discrepancy value observed across all bipartitions for the giventriple-chain (PhyloBayes tool).

chain ID	log Likelihood (harmonic mean, burnin-excl.)	chain combination ( <i>'triple'</i> chain)	<i>maxdiff</i> (< 0.3)
chain 18	948174.861012454	c04 - c18 - c20	0.186
chain 04	948217.993492174	c23 - c01 - c06	0.202933
chain 20	948376.710282837	c21 - c23 - c08	0.20787
chain 16	948469.642382507	c21 - c23 - c01	0.18833
chain 05	948525.74067471	c01 - c23 - c08	0.20787
chain 22	948678.821621205	c21 - c08 - c01	0.20787
chain 23	948708.71215524	c22 - c05 - c14	0.23647
chain 21	948752.989770425	c22 - c05 - c16	0.18653
chain 08	948757.764925626	c22 - c14 - c16	0.23647
chain 04	948779.209757328	c05 - c14 - c16	0.1621
chain 01	948865.845517544	all 25 chains	1

### Consensus network of single Bayesian topologies

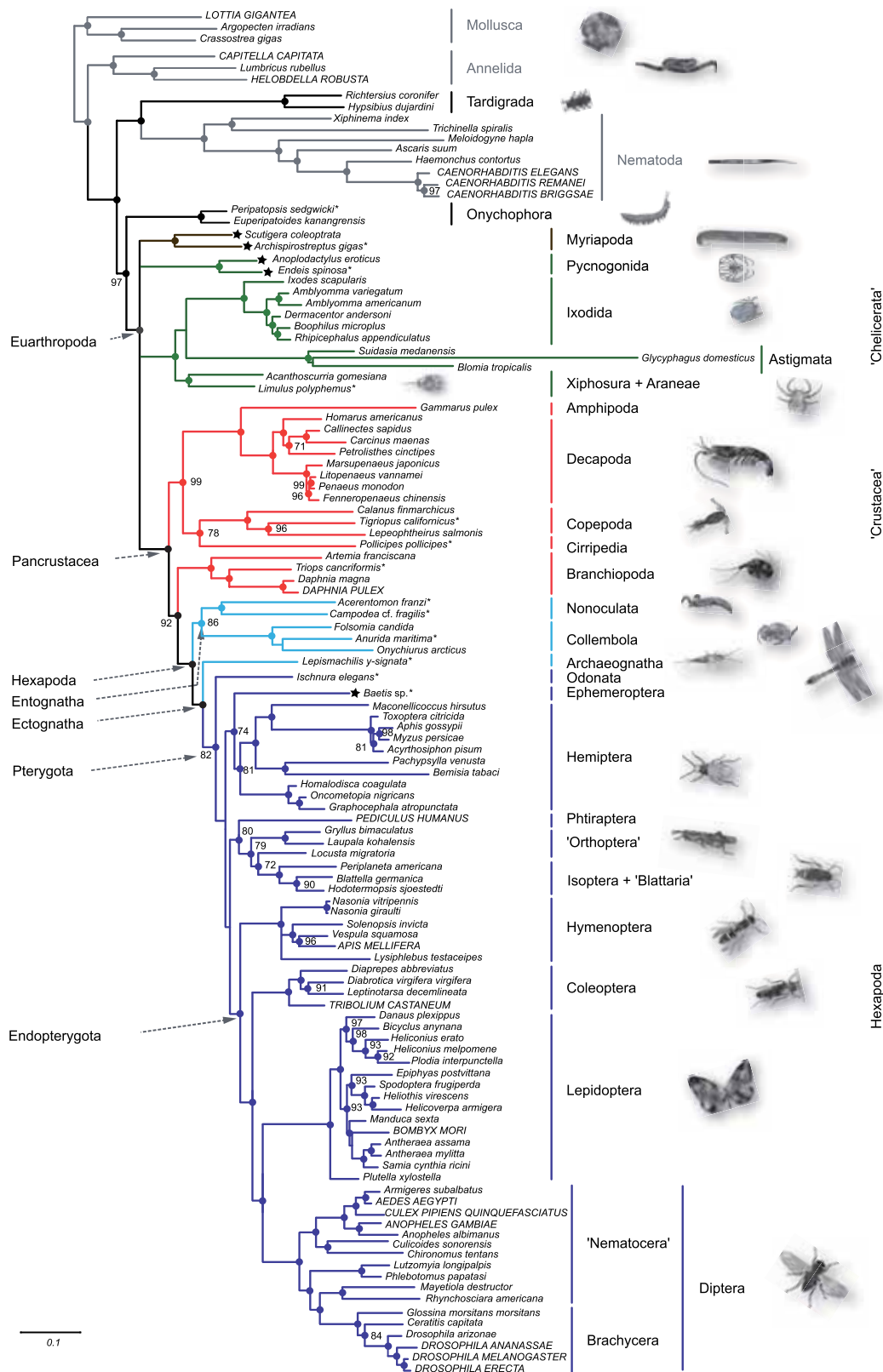
Due to differences between single topologies of the 25 PhyloBayes (Lartillot et al., 2008) chains, we computed a consensus network (Holland and Moulton, 2003) with SplitsTree 4.8 (Huson and Bryant, 2006). This is a method to identify contradictory signal that cannot be displayed with a simple majority rule consensus tree. To visualize conflicts and contradictory signal we chose a threshold of 0.01 and incorporated averaged edge weights.

## Results and Discussion

### The selected optimal subset (SOS)

Our selected optimal subset (SOS) includes 117 taxa with 101 euarthropods, two onychophorans, two tardigrades and twelve outgroup taxa (Supplementary table 1). The data set comprises 129 genes of which 32 genes coded for ribosomal proteins and 97 for non-ribosomal proteins (Supplementary table 2). The relative information content of genes ranges from 0.42 – 0.92, with an average of 0.7 (Supplementary tables 1, 2). The concatenated, masked alignment spans 37,476 amino acid positions (Supplementary figure 4). The relative informativeness rises fourfold from 0.10 (original data set) to 0.43 (SOS) (Supplementary figures 3-5). Matrix saturation (genes with a relative information content < 0.04 considered as missing) increases threefold from originally 17.6% to 62.3% in the SOS. Taxa in the SOS cover on average 84 genes (minimum 35, maximum 129). Each gene is, on average, present in 76 taxa (minimum 46, maximum 109 taxa per gene).

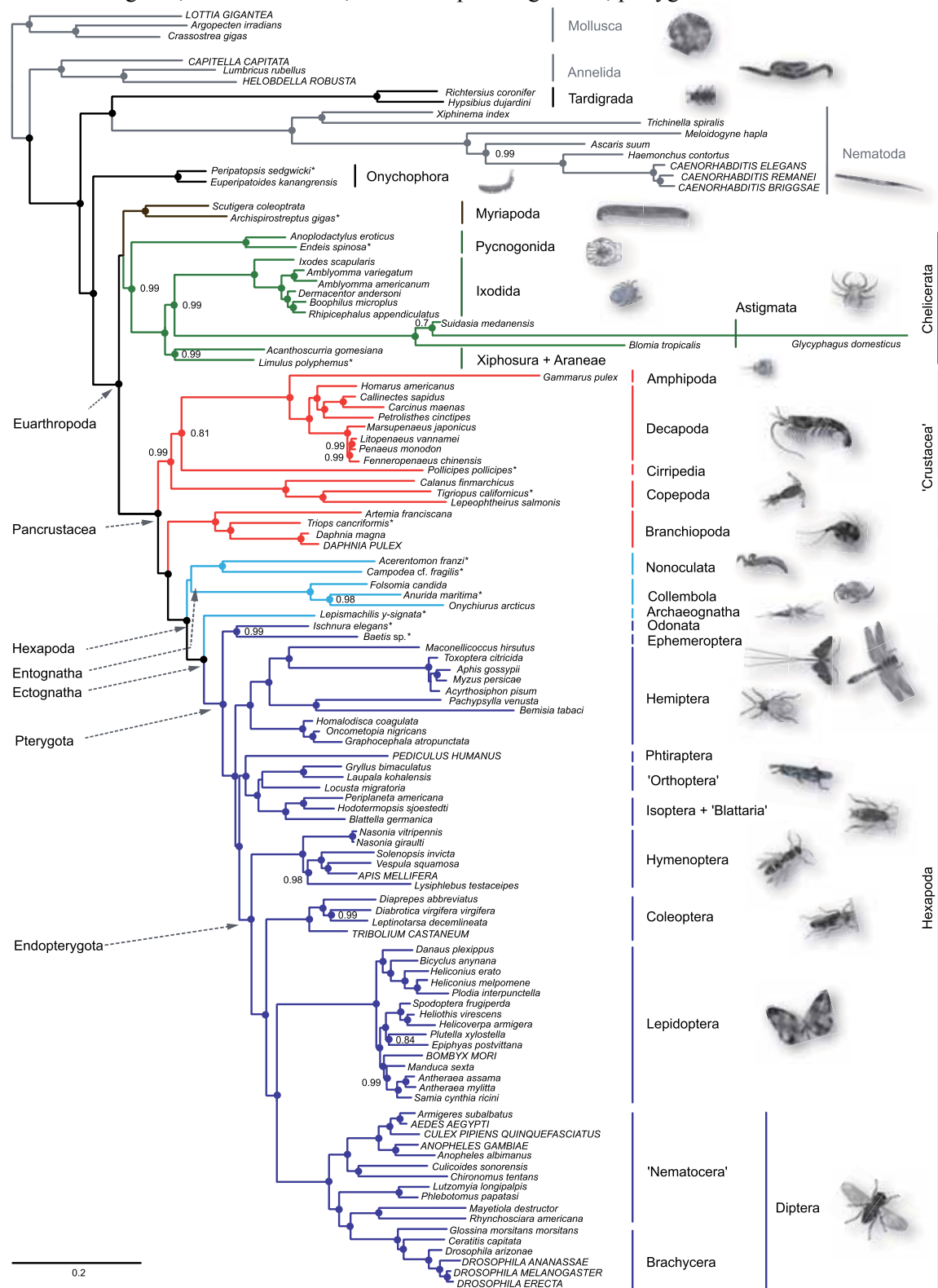
Maximum likelihood and Bayesian tree reconstruction of the SOS resolved arthropod relationships with several strongly supported nodes (figures 1, 2 and table 3). In contrast, the tree based on the original supermatrix is in many respects unresolved or shows low support values (Supplementary figure 6). This comparison suggests that the strategy to compute an SOS is successful, e.g. improves tree robustness and clades that are widely accepted in the literature (e.g. Hexapoda, Ectognatha, Endopterygota, Coleoptera, Lepidoptera etc.), which was not the case for the unreduced data set. Thus, the discussion of the phylogenetic relationships focuses on the SOS.



**Fig. 1.** – Phylogram of 117-taxon ML analysis.

RAxML tree (majority rule consensus) of the selected optimal subset (SOS), PROTMIX substitution model + WAG matrix. Support values are derived from 1,000 bootstrap replicates. Support values < 70: not shown, support values = 100: represented by a dot only. Quotation marks indicate non-monophyly. Asterisks (\*) indicate EST taxa contributed by the authors. Unstable taxa

(leaf stability index < 0.95) are marked by a star in front of the taxon name. Color code: molluscs, annelids and nematodes: lighter grey; tardigrades, onychophorans: black; myriapods: brown; chelicerates: green; crustaceans: red; basal hexapods: light blue; pterygote insects: dark blue.



**Fig. 2.** – Phylogram of 117-taxon Bayesian analysis.

Bayesian majority rule consensus tree of the selected optimal subset (SOS), 3 chains out of 25 chains, 20,000 cycles each, burn-in: 5,000 cycles. Posterior probabilities (pP) are estimated under the CAT mixture model. The majority rule consensus tree is based on the 'triple' (three chains) showing lowest maxdiff value (0.186) while each of these chains had the best harmonic mean of



the likelihood values (burn-in excluded) of all possible 'triple'-chain combinations. pP-values < 0.7: not shown, pP-values = 1.0: represented by a dot only. Quotation marks and color code as specified in figure 1; asterisks (\*) indicate EST taxa contributed by the authors.

**Table 3**

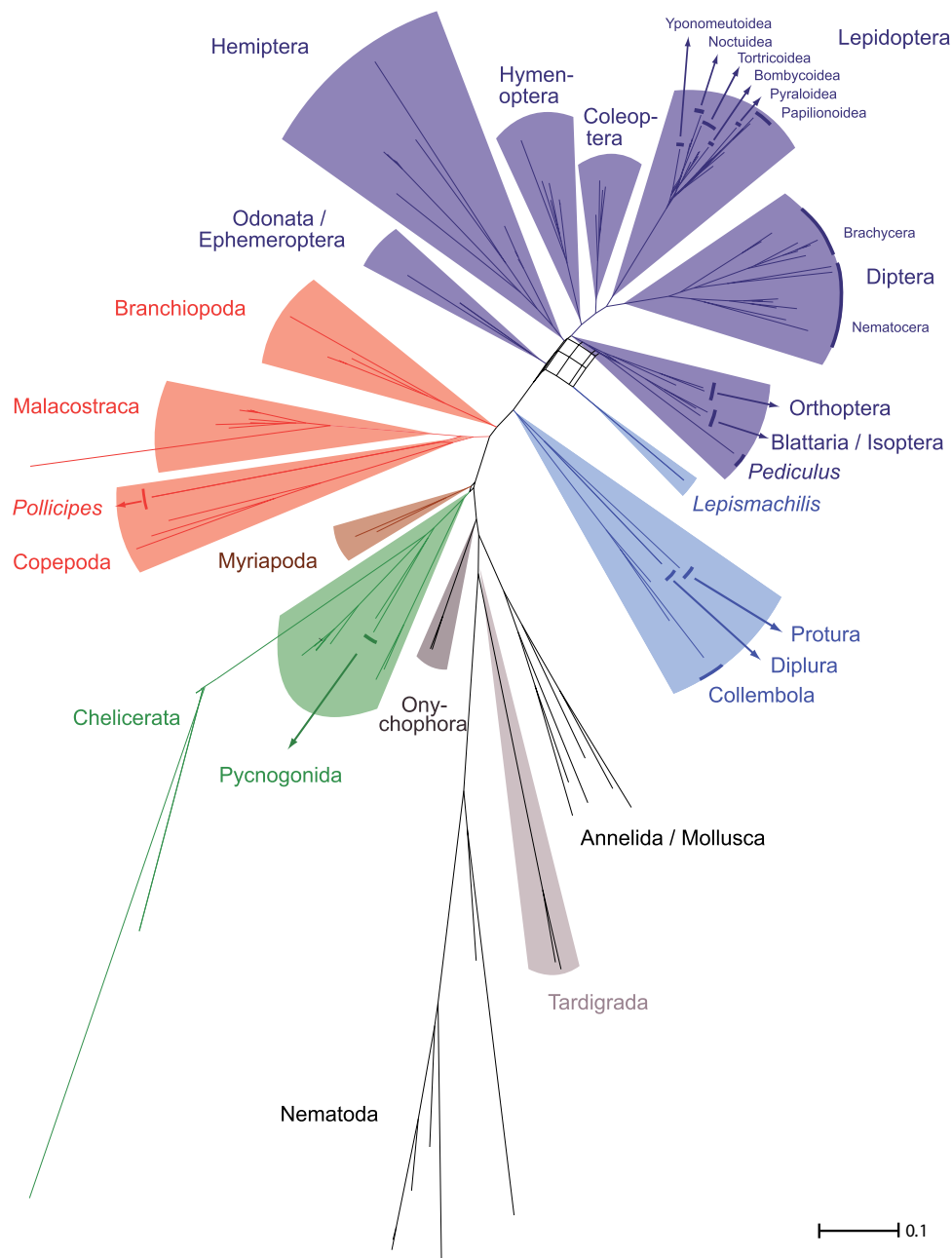
Selected clades and support values of maximum likelihood and Bayesian reconstructions inferred for the selected optimal subset (SOS).

Selected clades	Bootstrap Support [%]	posterior Probability
(Tardigrada,Nematoda)	100	1
(Onychophora,Euarthropoda)	97	1
((Tardigrada,Nematoda),(Onychophora,Euarthropoda))	100	1
Euarthropoda	100	1
Mandibulata	-	-
Myriochelata	-	0.57
Chelicerata	-	0.99
Euchelicerata	100	1
Pancrustacea	100	1
(Amphipoda,Decapoda)	100	1
(Copepoda,Cirripedia)	78	-
((Amphipoda,Decapoda),(Copepoda,Cirripedia))	99	-
((Amphipoda,Decapoda),Cirripedia)	-	0.81
((((Amphipoda,Decapoda),Cirripedia),Copepoda)	-	0.99
(Branchiopoda,Hexapoda)	92	1
Hexapoda	100	1
Enthognatha	86	0.5
(Collembola,(Protura,Diplura))	86	0.5
Nonoculata: (Protura,Diplura)	100	1
Ectognatha: (Archaeognatha,Pterygota)	100	1
Pterygota	82	1
Chiaistomyaria: (Odonata,(Ephemeroptera,Neoptera))	-	-
Paleoptera: (Odonata,Ehemeroptera)	-	0.99
Neoptera	-	1
(Ephemeroptera,Hemiptera)	74	-
Endopterygota	100	1
(Hymenoptera,remaining endopterygote clades)	100	1
(Coleoptera,(Lepidoptera,Diptera))	100	1
(Lepidoptera,Diptera)	100	1

### Incongruences in Bayesian analyses

The 25 Bayesian runs did not converge on a single topology (Methods, figure 3). Some clades, e.g. (Onychophora, Euarthropoda), Pancrustacea, Branchiopoda as a sister group to Hexapoda and Nonoculata (Protura, Diplura), emerged in all chains with maximal support. Other clades differed between consensus trees inferred from single chains. These incongruences were caused by unstable positions of few taxa (figure 3): **1)** Mandibulata (Myriapoda + Pancrustacea) were found maximally supported in consensus trees of two runs. Both runs show comparatively low harmonic means of likelihoods. In all other runs, myriapods clustered with chelicerates with negligible to moderate support (pP 0.52 – 0.89). **2)** The barnacle *Pollicipes* (Cirripedia, Crustacea) emerged as a sister group to copepods in only one run (pP 0.51). However, the alternative clade (*Pollicipes* + Malacostraca) (figure 2) showed a wide range from 0.56 – 0.96 posterior probability in other runs. **3)** The bristletail *Lepismachilis* (Archaeognatha) was inferred as a sister group to Blattaria +

Isoptera in several runs, showing moderate or low support (pP 0.52 – 0.82). Additionally *Pediculus* (Phthiraptera) emerged as a sister group to this clade (pP 1.0) in these runs. Likelihoods (harmonic means), however, were lower compared to runs used for our Bayesian consensus tree (figure 2), and results of these runs were rejected after a Bayes factor test (Kaas and Raftery, 1995; Nylander et al., 2004). 4) Among butterflies (Lepidoptera), five different topologies with distinctive clades were found. Differences occurred among Yponomeutoidea, Papilionoidea, Pyraloidea, Tortricoidea and Noctuoidea. Incongruent consensus trees might reflect different local optima despite extensive sampling.



**Fig. 3.** – Consensus network of all 25 PhyloBayes trees.

Consensus network of all 25 PhyloBayes chains of the selected optimal subset (SOS) was calculated with SplitsTree 4.8 and visualizes incongruences between 25 topologies (threshold = 0.01, averaged weights). The color code is specified in figure 1.

Are the enigmatic tardigrades and onychophorans arthropods sensu latu?

Chelicerates, myriapods, crustaceans and hexapods show highly derived differentiations of segments and segmental appendages (Edgecombe, 2009). Tardigrades and onychophorans display a mosaic of plesiomorphic and autapomorphic features of segmental differentiation. The evolution of the arthropod bauplan, as for example the evolution of segmentation, appendages and the central nervous system, can thus be understood only if the phylogenetic positions of tardigrades and onychophorans are resolved (see reviews in Budd and Telford, 2009; Edgecombe, 2009).

Tardigrades are tiny animals with morphological characters reminiscent of both Arthropoda and Cycloneuralia (the latter named for their circumpharyngeal nerve ring shared by Nematoda, Nematomorpha (horsehair worms, insect parasites), Priapulida (penis worms), Kinorhyncha (mud dragons), and Loricifera, see (Giribet, 2003; Edgecombe, 2009)). Arthropod-like characters include the segmented body, limbs, the presence of a peritrophic membrane and a ladder-like central nervous system (Giribet, 2003). In contrast, structures of mouth, pharynx, cuticle and sensory organs resemble those of Cycloneuralia (Giribet, 2003). Traditionally, tardigrades have been allied with arthropods, an assumption which has been corroborated by molecular studies based on ribosomal RNA (Mallatt et al., 2004). Such a clade Tardigrada + Onychophora + Euarthropoda (Panarthropoda) would be compatible with the hypothesis of an evolution of segmentation (including differentiation of the muscular tube, etc.), segmented appendages and a ladder-like central nervous system within this clade. Alternatively, a sister group relationship of tardigrades with Cycloneuralia (nematodes and allies) would imply either a very ancient evolution of a segmented body plan and a loss of these characteristics within derived Cycloneuralia (including a reversal to an undifferentiated muscular tube) or an independent evolution of segmental characters within Cycloneuralia. A robustly resolved position of tardigrades has a strong impact on our interpretation of the evolution of segmentation.

In our analyses, tardigrades (*Hypsibius* and *Richtersius*) emerge as a sister group of nematodes (bootstrap support 100%, posterior probability 1.0), which is in line with recent findings by Roeding et al. (2007), Lartillot and Philippe (2008) and Bleidorn et al. (2009). These studies had been based on different gene selections. In contrast, Dunn et al. (2008), applying the CAT model (Lartillot and Philippe, 2004) of amino acid evolution, found tardigrades either as a sister group of arthropods (including onychophorans), or applying

the WAG model, as a sister group of nematodes and nematomorphs, in both cases only weakly supported. Phylogenetic analyses based on morphological characters are similarly ambiguous and support contradicting results. Either panarthropods (including tardigrades, (Edgecombe, 2009)) are favored, or an unresolved clade (Tardigrada + Onychophora + Euarthropoda) is represented in Budd and Telford (2009), or tardigrades are positioned outside the (Onychophora + Euarthropoda) clade (Zantke et al., 2008). Currently, there is no conclusive hypothesis compatible with the contradicting morphological and molecular data about the position of tardigrades within the metazoan tree. This clearly impedes our understanding of the evolution of segmentation within Ecdysozoa.

Onychophorans strongly resemble arthropod-like animals with, for example, a reduction of locomotory cilia, a body cavity with a pericardial septum, a heart with ostia, segmental nephridia with sacculi, the presence of clawed ventral appendages and the absence of metameric larvae. Deviant from arthropods, onychophorans lack for example a complete disintegration of the muscular tube into segmentally arranged muscle systems, segmentally arranged sclerotized exoskeletal structures and a fully ganglionated organization of the central nervous system. Earlier morphological and molecular analyses have placed onychophorans as either a sister group to Tardigrada + Euarthropoda (Budd and Telford, 2009) or sister group to Euarthropoda (Roeding et al., 2007; Dunn et al., 2008; Edgecombe, 2009), thus leaving the position of onychophorans unresolved. Maximum likelihood (figure 1) and Bayesian (figure 2) analyses of our selected optimal subset (SOS) resolve the position of the onychophorans and show strong support for the clade Onychophora + Euarthropoda. A clade Onychophora + Euarthropoda is compatible with the view that fully differentiated segmentation, including ganglionization of the central nervous system evolved in a common stem-lineage of onychophorans and euarthropods. This view implies that onychophorans primarily lack many characteristics of the euarthropod body organization (Hou and Bergström, 1995; Edgecombe, 2009). The interpretation of the fossil record of “lobopodian”-grade organisms as possible stem group representatives of euarthropods is also compatible with this conclusion (Hou and Bergström, 1995).

Euarthropoda including pycnogonids favored over the “Cormogonida”

The monophyly of euarthropods is well established, whereas relationships within euarthropods, between myriapods, sea spiders, chelicerates, crustaceans and hexapods, are

problematic (compare results of Dunn et al., 2008; Regier et al., 2008; von Reumont et al., 2009; Regier et al., 2010).

Sea spiders (Pycnogonida) represent an extremely aberrant group of arthropods. Earlier morphological and molecular studies have placed sea spiders either as a sister group of Euchelicerata (Bourlat et al., 2008; Brenneis et al., 2008; Dunn et al., 2008) or considered them as the first branch of euarthropods ("Cormogonida" hypothesis, Zrzavy et al., 1998; Maxmen et al., 2005). While the position of sea spiders is not resolved in the ML tree (figure 1), the Bayesian tree (figure 2) shows monophyletic chelicerates including sea spiders with high support (posterior probability 0.99). This result corroborates other phylogenomic analyses (Dunn et al., 2008; Regier et al., 2010, but weakly supported) as well as hox gene and neuroanatomical studies (Jager et al., 2006; Brenneis et al., 2008), which demonstrated the homology of deutocerebral appendages of sea spiders and euchelicerates. It suggests that sea spiders should be included within chelicerates. Our results are inconclusive regarding the position of the pycnogonids, comparing the maximum likelihood and the Bayesian reconstruction, but the latter agrees with established "nonmolecular" data (Jager et al., 2006; Brenneis et al., 2008) that support pycnogonids as a sister group to Euchelicerata.

The position of Myriapoda cause problems to address Mandibulata vs. Myriochelata

Monophyly of mandibulate arthropods (Myriapoda + Crustacea + Hexapoda) has received substantial support from morphological studies (Richter, 2002; Harzsch et al., 2005; Harzsch, 2006; Scholtz and Edgecombe, 2006; Müller et al., 2007; Bäcker et al., 2008) and from some molecular analyses (Regier et al., 2008; Telford et al., 2008; Regier et al., 2010). Within mandibulates, two alternative clades, either Myriapoda + Hexapoda (Atelocerata (Heymons, 1901) or Tracheata (Pocock, 1893)) or Crustacea + Hexapoda (Pancrustacea (Zrzavy and Stys, 1997) or Tetraconata (Dohle, 2001)) have been proposed by Grimaldi (2009). Both hypotheses utilize the presence of complex character systems supporting each view (Harzsch, 2006; Bäcker et al., 2008; Mayer and Whittington, 2009). Molecular evidence, however, has recently accumulated for a clade Myriapoda + Chelicerata, coined Myriochelata (Pisani et al., 2004) or Paradoxopoda (Mallatt et al., 2004). This conflicts with the Mandibulata concept (Mallatt et al., 2004; Roeding et al., 2007; Dunn et al., 2008). At the same time, recent studies have demonstrated a high sensitivity of reconstructing Paradoxopoda with respect to gene choice, taxon sampling and outgroup selection (Bourlat et al., 2008; Philippe et al., 2009). The most recent study

addressing this issue was published by Regier et al. (2010) based on nuclear, mainly non-ribosomal protein coding genes which again supports Mandibulata. Ribosomal proteins, however, are hardly considered and this result should be interpreted with caution. Furthermore, there is little morphological data supporting a clade Paradoxopoda (Mayer and Whittington, 2009) in contrast to data supporting Mandibulata (Wägele, 1993; Harzsch, 2006; Bäcker et al., 2008). A clade Paradoxopoda would imply the independent evolution of the labium, the loss of the second pair of antennae and the independent evolution of ectodermal malphigian tubules in myriapods and hexapods.

In our analyses (including ribosomal and non-ribosomal single copy genes) the position of myriapods is not resolved. In the Bayesian tree, myriapods emerge as a sister group to chelicerates with low support. In the ML tree, relationships between myriapods, sea spiders, euchelicerates and pancrustaceans remain unresolved. The results of our phylogenomic analyses and rRNA based analyses (e.g. von Reumont et al., 2009) indicate that the unstable position of myriapods is not caused by a single myriapod taxon but probably is related to a systematic phenomenon of myriapod molecular evolution. To resolve the myriapod position in the arthropod tree, we therefore need to better understand heterogeneity of substitutional processes among arthropods and to include all myriapod groups in phylogenomic analyses.

#### Pancrustacea with Branchiopoda as a sister group to Hexapoda

Our data support a clade Crustacea + Hexapoda (Pancrustacea, 100% bootstrap support and 1.0 posterior probability). Within crustaceans, relationships are still far from being resolved. Representatives of important crustacean groups are still not covered by EST data. Only few published non-malacostracan EST projects exist (Branchiopoda, Copepoda and Cirripedia, presented in this study) (Stillman et al., 2008). Therefore, discussing the sister group of hexapods requires caution, and further EST data for representatives of major crustacean groups (e.g. Remipedia, Leptostraca) are required.

In rRNA based studies, copepods (Cyclopidae) were found to be a sister group to hexapods (Mallatt and Giribet, 2006; von Reumont et al., 2009; Mallatt et al., 2010). In our analyses Branchiopoda consistently emerge as a sister group to Hexapoda (1.0 posterior probability in the Bayesian approach and moderately supported 92% bootstrap support in ML analyses). This corroborates results of other single- and multi-gene analyses (Regier et al., 2005; Dunn et al., 2008; Philippe et al., 2009; Mallatt et al., 2010). This well-supported clade Branchiopoda + Hexapoda conflicts with described potential synapomorphies of

Malacostraca and Hexapoda (Harzsch, 2006), e.g. the presence of a third neuropil and chiasmata of the lateral eyes. Ertas et al. (2009) suggest a close relationship of Remipedia and Hexapoda based on hemocyanin. This result is underpinned by neuroanatomical data (Fanenbruck et al., 2004; Fanenbruck and Harzsch, 2005). Regier et al. (2010) inferred a clade “Xenocarida” with Remipedia + Cephalocarida as a sister group to Hexapoda, with low support at the amino acid level and high support at nucleotide level. Remipedia as the sister group to Cephalocarida is contradicted by new data on Remipedia larvae (Koenemann et al., 2007; Koenemann et al., 2009). The incongruence between molecular and morphological results concerning the sister group relationship of hexapods cannot be resolved yet. Careful analyses of signal quality in molecular and morphological data are still required, along with more molecular data on Remipedia and Cephalocarida.

#### Monophyletic Hexapoda, Entognatha and Ectognatha

Based on morphological analyses, hexapods are assumed to be monophyletic (Dohle, 2001; Bitsch and Bitsch, 2004; Harzsch et al., 2005; Harzsch, 2006; Ungerer and Scholtz, 2008). The monophyly of ectognathous hexapods (Archaeognatha + pterygote insects, see Hennig (1981) and Kristensen (1991)) seems well founded by single-gene analyses (Kjer et al., 2006; Misof et al., 2007; von Reumont et al., 2009), is supported by nuclear protein coding genes (Regier et al., 2010) and also corroborated by our phylogenomic data; this clade “has likewise never been seriously challenged” (Grimaldi, 2009).

In contrast, the monophyly of entognathous hexapods (Protura, Diplura and Collembola) is generally ambiguous (see review in Grimaldi, 2009). The interpretation of character states within entognathous hexapods is difficult because of extreme adaptations to subterranean or cryptic habitats. The presence of many plesiomorphic character states (e.g. presence of fully muscled antennae, abdominal appendages, anameric development (Protura), unsegmented tarsi) gives them an important role in understanding the evolution of hexapods. Our Bayesian and ML analyses recovered Entognatha as a monophyletic group, albeit weakly supported. Within Entognatha, we obtain strong support for a sister group relationship of Protura and Diplura, a clade coined Nonoculata (Luan et al., 2005). This corroborates recent single gene analyses (Dell'Ampio et al., 2009; von Reumont et al., 2009; Mallatt et al., 2010). Morphological evidence for this clade is still ambiguous (Szucsich and Pass, 2008). Our results disagree with inferred relationships of primary wingless hexapods based on mitochondrial data (Nardi et al., 2003; Carapelli et al., 2005; Carapelli et al., 2007). Those authors proposed the polyphyly of hexapods with a

placement of springtails (Collembola) as a sister group to other pancrustacean taxa, implying that features of the hexapod bauplan evolved at least twice. Reanalyses of these mitochondrial data (Delsuc et al., 2003) yielded monophyletic hexapods (although weakly supported). Those analyses, however, never included proturans. Also in recent studies both Protura and Diplura (e.g. Timmermans et al., 2008; Aleshin et al., 2009), or at least Protura, are missing (Regier et al., 2008; Regier et al., 2010). Including these orders is indispensable to infer deep hexapod relationships. Our analyses based on much more extensive phylogenomic data, including all orders of monocondyl, primary wingless hexapods, yielded strong support for monophyletic hexapods. We conclude that hexapods are monophyletic and that the distinctive bauplan evolved only once.

Relationships among pterygote insects are still disputed. A puzzling problem is the early evolution of winged insects (Whitfield and Kjer, 2008). Mayflies, dragonflies and neopterous winged insects appear early in the fossil record. Morphological and molecular analyses either support a clade (Odonata (Ephemeroptera + Neoptera)) coined “Chiastomyaria” (Boudreaux, 1979; Kjer, 2004), or “Metapterygota” (Ephemeroptera (Odonata + Neoptera)), see Börner (1909) and Zhang et al. (2008), or “Palaeoptera” ((Odonata + Ephemeroptera) Neoptera), see Hennig (1981) and Kukalová-Peck (1983). Most molecular analyses support either a “Chiastomyaria” or “Palaeoptera” clade (see discussion in Simon et al., 2009). A possible explanation for the difficult-to-resolve relationships is an ‘explosive radiation’ once flight evolved (Whitfield and Kjer, 2008). Our phylogenomic data are inconclusive in ML tree reconstructions, but strongly support “Palaeoptera” in Bayesian analyses. Convincing morphological synapomorphies for Paleoptera and Neoptera are lacking.

Within neopterous insects, relationships among endopterygote insects are a major focus of scientific activity. For example, it is unclear whether beetles + neuropteridans (Neuropteroidea) branch off first or whether hymenopterans are the sister group to all other endopterygote insects (Kristensen, 1999; Kukalová-Peck and Lawrence, 2004; Beutel and Pohl, 2006; Wiegmann et al., 2009). Our analyses strongly support most orders of Endopterygota (figures 1, 2). Several of these clades corroborate previous results based on single nuclear genes (von Reumont et al., 2009). Our phylogenomic approach also unambiguously supports hymenopterans as the sister group to all other endopterygote insects and corroborates previous studies (e.g. Savard et al., 2006; Simon et al., 2009; Wiegmann et al., 2009), in contrast to conclusions based on complete mitochondrial genomes (Castro and Dowton, 2005). This result will be extremely important in



interpreting and understanding early extinct endopterygote insects and the evolution of this most species-rich group of arthropods.

## **Conclusions**

We show that phylogenomic studies, although raising hope to reach a resolved arthropod tree, still face challenges in interpreting the strength and quality of the phylogenetic signal. We also illustrate unresolved incongruences between morphological and molecular analyses. This, in our opinion, should challenge systematists of every camp to present the strength, quality and deficiencies of their evidence, and work towards resolving outstanding issues.

## **Supplementary Material**

The Supplementary Material containing Supplementary tables 1-2, Supplementary figures 1-6 and Supplementary literature are available at Molecular Biology and Evolution online (<http://www.mbe.oxfordjournals.org/>).

## **Funding**

This is a collaboration of the 'Arthropod Network' within the DFG Priority Program 1174 "Deep Metazoan Phylogeny" (<http://www.deep-phylogeny.org>). B.M. and K.M. were supported by the DFG grant MI 649/62, J.W.W., P.K. and B.M.v.R. were supported by grants WA530/33 and WA530/34. H.H. and S.Si. were funded by the DFG grant HA 1947/5, F.R. and T.B. by the DFG grant Bu956/9. S.St. was granted by the DFG project HA 1628/8, A.v.H. and I.E. were supported by the WWTF. G.P. was funded by the Austrian Science Foundation (FWF) grant P 20497-B17; the SuGI-project (S.B. and V.A.) is supported by the BMBF, Germany.

## **Acknowledgments**

We acknowledge Nikola Szucsich, Daniela Bartel, Johannes Dambach, Markus Pennerstorfer, Erich Eder, Francisco Javier Cristobo, Hilke Ruhberg and Sara Khadjeh for sampling and delivering specimens or tissues. We are also grateful to Michael Kube and Richard Reinhardt (MPIMG, Berlin, Germany) for the construction and sequencing of cDNA libraries and for submitting assembled contigs. We thank Benjamin Meyer, Nikola Szucsich, Janus Borner, Roman Stocsits and Stefan Grunewald for discussion and help

with software applications, and Alexandros Stamatakis for RAxML support. Special thanks go to Katharina Misof and Michael Stachowitsch for linguistic help. We also gratefully thank Barbara Holland and anonymous reviewers for providing constructive comments that improved this manuscript.

The data matrices are available from Treebase ([www.treebase.org](http://www.treebase.org)) or from <http://www.zfmk.de/web/Forschung/Molekularlabor/Datenstze/index.en.html>.

### **Author's contributions**

F.R. and T.B. delivered EST data for onychophorans, myriapods, pycnogonids and euchelicerates. B.M.v.R. and J.W.W. supplied ESTs for all new crustacean species. K.M. and B.M. provided ESTs for four hexapod species, M.W. and G.P. enabled the proturan EST project to be conducted. S.Si. and H.H. delivered EST data for two pterygote insects. Processing of EST data and orthologous gene prediction were performed by S.St. and I.E. Alignment masking and new reduction heuristics were developed by B.M. K.M., B.M., B.M.v.R. designed the study, and analyses were conducted by K.M., B.M.v.R., S.Si. and B.M. P.K. provided Perl-Scripts and was involved in development of reduction heuristics, S.B. and V.A. enabled all Bayesian analyses with technical and computational support. The manuscript was written by B.M., K.M., B.M.v.R., T.B., S.Si. and S.St. with comments and revisions from J.W.W., G.P., I.E., S.B., V.A. and A.v.H. All authors read and approved the final manuscript.

### **References**

- Aleshin VV, Mikhailov KV, Konstantinova AV, et al. 2009. On the Phylogenetic Position of Insects in the Pancrustacea Clade. *Molecular Biology*. 43: 804-818.
- Bäcker H, Fanenbruck M, Wägele JW 2008. A forgotten homology supporting the monophyly of Tracheata: The subcoxa of insects and myriapods re-visited. *Zool Anz*. 247: 185-207.
- Beutel RG, Pohl H 2006. Endopterygote systematics - where do we stand and what is the goal (Hexapoda, Arthropoda)? *Systematic Entomology*. 31: 202-219.
- Bitsch C, Bitsch J 2004. Phylogenetic relationships of basal hexapods among the mandibulate arthropods: a cladistic analysis based on comparative morphological characters. *Zoologica Scripta*. 33: 511-550.
- Bleidorn C, Podsiadlowski L, Zhong M, et al. 2009. On the phylogenetic position of Myzostomida: can 77 genes get it wrong? *Bmc Evolutionary Biology*. 9: -.
- Börner C 1909. Neue Homologien zwischen Crustaceen und Hexapoden. Die Beissmandibel der Insecten und ihre phylogenetische Bedeutung. *Archi-und Metapterygota*. *Zool Anz*. 34: 100-125.

- Boudreaux BH 1979. Arthropod phylogeny: with special reference to insects John Wiley & Sons Inc.
- Bourlat SJ, Juliusdottir T, Lowe CJ, et al. 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature*. 444: 85-88.
- Bourlat SJ, Nielsen C, Economou AD, Telford MJ 2008. Testing the new animal phylogeny: A phylum level molecular analysis of the animal kingdom. *Molecular Phylogenetics and Evolution*. 49: 23-31.
- Brenneis G, Ungerer P, Scholtz G 2008. The chelifores of sea spiders (Arthropoda, Pycnogonida) are the appendages of the deutocerebral segment. *Evolution & Development*. 10: 717-724.
- Budd GE, Telford MJ 2009. The origin and evolution of arthropods. *Nature*. 457: 812-817.
- Carapelli A, Lio P, Nardi F, van der Wath E, Frati F 2007. Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea. *Bmc Evolutionary Biology*. 7: -.
- Carapelli A, Nardi F, Dallai R, et al. 2005. Relationships between hexapods and crustaceans based on four mitochondrial genes. In: *Crustacean and Arthropod Relationships*, pp. 295-306. CRC Press.
- Castresana J 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 17: 540-552.
- Castro LR, Dowton M 2005. The position of the Hymenoptera within the Holometabola as inferred from the mitochondrial genome of *Perga condei* (Hymenoptera : Symphyta : Pergidae). *Molecular Phylogenetics and Evolution*. 34: 469-479.
- Chou HH, Holmes MH 2001. DNA sequence quality trimming and vector removal. *Bioinformatics*. 17: 1093-1104.
- Dell'Ampio E, Szucsich NU, Carapelli A, et al. 2009. Testing for misleading effects in the phylogenetic reconstruction of ancient lineages of hexapods: influence of character dependence and character choice in analyses of 28S rRNA sequences. *Zool Scr*. 38: 155-170.
- Delsuc F, Phillips MJ, Penny D 2003. Comment on "Hexapod origins: Monophyletic or paraphyletic?" *Science*. 301: 1482-1483.
- Delsuc F, Tsagkogeorga G, Lartillot N, Philippe H 2008. Additional molecular support for the new chordate phylogeny. *Genesis*. 46: 592-604.
- Dohle W 2001. Are the insects terrestrial crustaceans? A discussion of some new facts and arguments and the proposal of the proper name ``Tetraconata" for the monophyletic unit Crustacea + Hexapoda. *Ann Soc Entomol Fr (New Series)*. 37: 85-103.
- Dunn CW, Hejnol A, Matus DQ, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*. 452: 745-749.
- Ebersberger I, Strauss S, von Haeseler A 2009. HaMStR: Profile Hidden Markov Model Based Serach for Orthologs in ESTs. submitted.
- Edgecombe GD 2009. Palaeontological and molecular evidence linking arthropods, onychophorans, and other Ecdysozoa. *Evo Edu Outreach*. 2: 178-190.
- Ertas B, von Reumont BM, Wagele JW, Misof B, Burmester T 2009. Hemocyanin suggests a close relationship of Remipedia and Hexapoda. *Mol Biol Evol*. 26: 2711-2718.
- Fanenbruck M, Harzsch S 2005. A brain atlas of *Godzilliognomus frondosus* Yager, 1989 (Remipedia, Godzilliidae) and comparison with the brain of *Speleonectes tulumensis* Yager, 1987 (Remipedia, Speleonectidae): implications for arthropod relationships. *Arthropod Structure & Development*. 34: 343-378.
- Fanenbruck M, Harzsch S, Wagele JW 2004. The brain of the Remipedia (Crustacea) and an alternative hypothesis on their phylogenetic relationships. *Proc Natl Acad Sci U S A*. 101: 3868-3873.

- Giribet G 2003. Molecules, development and fossils in the study of metazoan evolution; Articulata versus Ecdysozoa revisited. *Zoology (Jena)*. 106: 303-326.
- Grimaldi DA 2009. 400 million years on six legs: On the origin and early evolution of Hexapoda. *Arthropod Struct Dev*.
- Hartmann S, Vision TJ 2008. Using ESTs for phylogenomics: can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evol Biol*. 8: 95.
- Harzsch S 2006. Neurophylogeny: Architecture of the nervous system and a fresh view on arthropod phylogeny. *Integr Comp Biol*. 46: 162-194.
- Harzsch S, Müller CHG, Wolf H 2005. From variable to constant cell numbers: cellular characteristics of the arthropod nervous system argue against a sister-group relationship of Chelicerata and "Myriapoda" but favour the Mandibulata concept. *Dev Genes Evol*. 215: 53-68.
- Hejnol A, Obst M, Stamatakis A, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc Biol Sci*. 276: 4261-4270.
- Hennig W 1981. *Insect Phylogeny* John Wiley & Sons, New York.
- Heymonds R 1901. Die Entwicklungsgeschichte der Scolopender. *Zoologica*. H. 33: 1-244, Taf. I-VIII.
- Holland B, Moulton V (2003) Consensus networks: A method for visualising incompatibilities in collections of trees. In: *Workshop on Algorithms in Bioinformatics (WABI) 2812*, (ed. Proceedings. LNBI), pp. 165-176. Springer, Berlin / Heidelberg.
- Holmes DS, Bonner J 1973. Preparation, Molecular-Weight, Base Composition, and Secondary Structure of Giant Nuclear Ribonucleic-Acid. *Biochemistry*. 12: 2330-2338.
- Hou X, Bergström J 1995. Cambrian lobopodians -ancestors of extant onychophorans? *Zool J Linn Soc*. 114: 3-19.
- Huson DH, Bryant D 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*. 23: 254-267.
- Jager M, Murienne J, Clabaut C, et al. 2006. Homology of arthropod anterior appendages revealed by Hox gene expression in a sea spider. *Nature*. 441: 506-508.
- Kaas RE, Raftery AE 1995. Bayes Factors. *Journal of the American Statistical Association*. 90: 773-795.
- Katoh K, Toh H 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform*. 9: 286-298.
- Kjer KM 2004. Aligned 18S and insect phylogeny. *Syst Biol*. 53: 506-514.
- Kjer KM, Carle FL, Litman J, Ware J 2006. A Molecular Phylogeny of Hexapoda. *Arthropod Systematics & Phylogeny*. 64: 35-44.
- Koenemann S, Olesen J, Alwes F, et al. 2009. The post-embryonic development of Remipedia (Crustacea)-additional results and new insights (vol 219, pg 131, 2009). *Development Genes and Evolution*. 219: 217-217.
- Koenemann S, Schram FR, Bloechl A, et al. 2007. Post-embryonic development of remipede crustaceans (vol 9, pg 117, 2007). *Evolution & Development*. 9: 306-306.
- Kristensen NP (1991) Phylogeny of extant hexapods. In: *The Insects of Australia: A Textbook for Students and Research Workers* (ed. Naumann ID, Lawrence, J.F., Nielsen, E.S., Spradberry, J.P., Taylor, R.W., Whitten, M.J., Littlejohn, M.J.). CSIRO, Melbourne Univ. Press, Melbourne.
- Kristensen NP 1999. Phylogeny of endopterygote insects, the most successful lineage of living organisms. *Eur J Entomol*. 96: 237-253.
- Kück P, Meusemann K, Dambach J, et al. 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. *Front Zool*. forthcoming.

- Kukalová-Peck J 1983. Origin of the insect wing and wing articulation from the arthropodan leg. *Can J Zool.* 61: 1618-1669.
- Kukalová-Peck J, Lawrence JF 2004. Relationships among coleopteran suborders and major endoneopteran lineages: Evidence from hind wing characters. *Eur J Entomol.* 101: 95-144.
- Lartillot N, Blanquart S, T. L 2008. PhyloBayes 2.3 -a Bayesian software for phylogenetic reconstruction using mixture models (2.3c, current versions available from <http://www.phylobayes.org> edn), University of Montreal.
- Lartillot N, Philippe H 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution.* 21: 1095-1109.
- Lartillot N, Philippe He 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philos Trans R Soc Lond, B, Biol Sci.* 363: 1463-1472.
- Luan Y-x, Mallatt JM, Xie R-d, Yang Y-m, Yin W-y 2005. The phylogenetic positions of three basal-hexapod groups (Protura, Diplura, and Collembola) based on on ribosomal RNA gene sequences. *Mol Biol Evol.* 22: 1579-1592.
- Mallatt J, Craig CW, Yoder MJ 2010. Nearly complete rRNA genes assembled from across the metazoan animals: effects of more taxa, a structure-based alignment, and paired-sites evolutionary models on phylogeny reconstruction. *Mol Phylogenet Evol.* 55: 1-17.
- Mallatt J, Giribet G 2006. Further use of nearly complete 28S and 18S rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. *Mol Phylogenet Evol.* 40: 772-794.
- Mallatt JM, Garey JR, Shultz JW 2004. Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Mol Phylogenet Evol.* 31: 178-191.
- Maxmen A, Browne WE, Martindale MQ, Giribet G 2005. Neuroanatomy of sea spiders implies an appendicular origin of the protocerebral segment. *Nature.* 437: 1144-1148.
- Mayer G, Whittington PM 2009. Velvet worm development links myriapods with chelicerates. *Proc Biol Sci.* 276: 3571-3579.
- Misof B, Misof K 2009. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Syst Biol.* 58: syp006.
- Misof B, Niehuis O, Bischoff I, et al. 2007. Towards an 18S phylogeny of hexapods: Accounting for group-specific character covariance in optimized mixed nucleotide/doublet models. *Zoology (Jena).* 110: 409-429.
- Müller CHG, Sombke A, Rosenberg J 2007. The fine structure of the eyes of some bristly millipedes (Penicillata, Diplopoda): Additional support for the homology of mandibulate ommatidia. *Arthropod Struct Dev.* 36: 463-476.
- Nardi F, Spinsanti G, Boore JL, et al. 2003. Hexapod origins: monophyletic or paraphyletic? *Science.* 299: 1887-1889.
- Nieselt-Struwe K, von Haeseler A 2001. Quartet-mapping, a generalization of the likelihood-mapping procedure. *Molecular Biology and Evolution.* 18: 1204-1219.
- Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JeL 2004. Bayesian phylogenetic analysis of combined data. *Syst Biol.* 53: 47-67.
- Ott M, Zola J, Aluru S, Stamatakis A (2007) Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L.
- Pertea G, Huang X, Liang F, et al. 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics.* 19: 651-652.

- Philippe H, Derelle R, Lopez P, et al. 2009. Phylogenomics revives traditional views on deep animal relationships. *Curr Biol.* 19: 706-712.
- Pisani D, Poling L, Lyons-Weiler M, Hedges SB 2004. The colonization of land by animals: molecular phylogeny and divergence times among arthropods. *BMC Biol.* 2: 1.
- Pocock RI 1893. On the classification of the tracheate Arthropoda. *Zool Anz.* 16: 271-275.
- Regier JC, Shultz JW, Ganley ARD, et al. 2008. Resolving Arthropod Phylogeny: Exploring Phylogenetic Signal within 41 kb of Protein-Coding Nuclear Gene Sequence. *Systematic Biology.* 57: 920-938.
- Regier JC, Shultz JW, Kambic RE 2005. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Proc Biol Sci.* 272: 395-401.
- Regier JC, Shultz JW, Zwick A, et al. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature.* 463: 1079-1098.
- Richter S 2002. The Tetraconata concept: hexapodcrustacean relationships and the phylogeny of Crustacea. *Org Divers Evol.* 2: 217-237.
- Roeding F, Hagner-Holler S, Ruhberg H, et al. 2007. EST sequencing of Onychophora and phylogenomic analysis of Metazoa. *Mol Phylogenet Evol.* 45: 942-951.
- Sanderson MJ 2007. Construction and annotation of large phylogenetic trees. *Aust Syst Bot.* 20: 287-301.
- Savard J, Tautz D, Richards S, et al. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. *Genome Res.* 16: 1334-1338.
- Scholtz G, Edgecombe GD 2006. The evolution of arthropod heads: reconciling morphological, developmental and palaeontological evidence. *Development Genes and Evolution.* 216: 395-415.
- Simon S, Strauss S, von Haeseler A, Hadrys H 2009. A phylogenomic approach to resolve the basal pterygote divergence. *Mol Biol Evol.* 26: 2719-2730.
- Smith SA, Dunn CW 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics.* 24: 715-716.
- Stamatakis A (2006a) Phylogenetic models of rate heterogeneity: A high performance computing perspective.
- Stamatakis A 2006b. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 22: 2688-2690.
- Stillman JH, Colbourne JK, Lee CE, et al. 2008. Recent advances in crustacean genomics. *Integr Comp Biol.* 48: 852-868.
- Szucsich NU, Pass Gu 2008. Incongruent phylogenetic hypotheses and character conflicts in morphology: The root and early branches of the hexapodan tree. *Mitt Dtsch Ges Allg Angew Entomol.* 16: 415-429.
- Telford MJ, Bourlat SJ, Economou A, Papillon D, Rota-Stabelli O 2008. The evolution of the Ecdysozoa. *Philos Trans R Soc Lond B Biol Sci.* 363: 1529-1537.
- Thorley JL, Wilkinson M 1999. Testing the phylogenetic stability of early tetrapods. *J Theor Biol.* 200: 343-344.
- Timmermans MJ, Roelofs D, Marien J, van Straalen NM 2008. Revealing pancrustacean relationships: phylogenetic analysis of ribosomal protein genes places Collembola (springtails) in a monophyletic Hexapoda and reinforces the discrepancy between mitochondrial and nuclear DNA markers. *BMC Evol Biol.* 8: 83.
- Ungerer P, Scholtz G 2008. Filling the gap between identified neuroblasts and neurons in crustaceans adds new support for Tetraconata. *Proc Biol Sci.* 275: 369-376.
- von Reumont BM, Meusemann K, Szucsich NU, et al. 2009. Can comprehensive background knowledge be incorporated into substitution models to improve

- phylogenetic analyses? A case study on major arthropod relationships. *BMC Evol Biol.* 9: 119.
- Wägele JW 1993. Rejection of the 'Uniramia' hypothesis and implications of the Mandibulata concept. *Zoologische Jahrbücher. Abteilung für Systematik, Ökologie und Geographie der Tiere.* 120: 253-288.
- Whelan S, Goldman N 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol.* 18: 691-699.
- Whitfield JB, Kjer KM 2008. Ancient rapid radiations of insects: challenges for phylogenetic analysis. *Annu Rev Entomol.* 53: 449-472.
- Wiegmann BM, Trautwein MD, Kim JW, et al. 2009. Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. *BMC Biol.* 7: 34.
- Wiens JJ, Moen DS 2008. Missing data and the accuracy of Bayesian phylogenetics. *J Syst Evol.* 46: 307-314.
- Zantke J, Wolff C, Scholtz G 2008. Three-dimensional reconstruction of the central nervous system of *Macrobiotus hufelandi* (Eutardigrada, Parachela): implications for the phylogenetic position of Tardigrada. *Zoomorphology.* 127: 21-36.
- Zhang J, Zhou C, Gai Y, Song D, Zhou K 2008. The complete mitochondrial genome of *Parafronurus youi* (Insecta: Ephemeroptera) and phylogenetic position of the Ephemeroptera. *Gene.* 424: 18-24.
- Zrzavy J, Mihulka S, Kepka P, Bezdek A, Tietz D 1998. Phylogeny of the Metazoa based on morphological and 18S ribosomal DNA evidence. *Cladistics-the International Journal of the Willi Hennig Society.* 14: 249-285.
- Zrzavy J, Stys P 1997. The basic body plan of arthropods: insights from evolutionary morphology and developmental biology. *J Evol Biol.* 10: 653-367.

**The mitochondrial genome of two palaeopterous representatives:  
*Baetis* sp. (Ephemeroptera) and *Boyeria irene* (Odonata) –  
a mitogenomic approach to resolve the Palaeoptera problem**

**Sabrina Simon<sup>1,\*</sup> and Heike Hadrys<sup>1,2</sup>**

<sup>1</sup> *ITZ, Ecology & Evolution, Stiftung Tierärztliche Hochschule Hannover, D-30559 Hannover, Germany*

<sup>2</sup> *Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut 06520, USA*

\*Corresponding author

This is the author's version of a work in preparation for submission to *BMC Genome*.



## Abstract

### Background

The phylogenetic positions of the basal winged insect orders (Palaeoptera: Ephemeroptera and Odonata) remain controversial or unresolved, but are extremely important to understand the macro-evolutionary transition from wingless to winged insects. In recent years, several genes, as well as different molecular marker systems have been used to clarify this major question in insect phylogeny. The phylogenetic analyses have shown the incongruence among inferred phylogenetic relationships using different molecular datasets. Based on nuclear data (rRNA genes) as well as on phylogenomic data, Odonata represents the earliest branching pterygote lineage while mitogenomic data supports Ephemeroptera as the basal pterygote order. In this study, we re-evaluate the Palaeoptera problem by conducting a mitogenomic approach with an improved taxon-sampling of basal pterygote orders.

### Results

We have sequenced the mitochondrial genome of the mayfly *Baetis* sp. and the dragonfly *Boyeria irene* and performed phylogenetic analyses for both species. Our results based on the protein-coding genes of a subset of available complete mitochondrial genome sequences of insects highlights the lack of phylogenetic signal for the deep nodes among insects. Furthermore we show the sensitivity of mitogenomic analyses for inferred deep insect divergences to additional sequence information.

### Conclusions

The availability of complete mitochondrial genomes across insects is a powerful source for comparative mitogenomics and phylogenetic studies on different taxonomic scales. However, the Palaeoptera problem could maybe only be resolved when appropriate evolutionary models are used in conjunction with additional sequence information derived from improved taxon sampling and/or inclusion of mitochondrial rRNAs/tRNAs in the phylogenetic analyses. Our analyses underline the caution which is needed when applying mitochondrial genome sequences to unravel the deep evolutionary history among basal insects.

## Background

The mitochondrial genome of the majority of metazoans is a circular molecule ranging in size from ~14 kb to ~16 kb and contains 37 genes: 13 protein coding genes (PCGs), two ribosomal RNAs (16S and 12S rRNA) and 22 transfer-RNAs (tRNAs), although some deviations in size and gene content are observed in non-bilaterian mitogenomes; see [1]. The transcription of all encoded genes can be distributed between the two strands [2]. The larger non-coding control region (AT-rich) typically occurs in mitochondrial genomes of hexapods [3] and includes the origin of replication [4]. The mitochondrial DNA uses a specific code for protein translation which differs from the code of the nuclear genome [5]. Incomplete stop-codons like T or TA occur in addition to TAA triplets, and are completed by polyadenylation of the messenger RNA. Mitochondrial genomes are further characterized by short overlaps of adjacent genes that are encoded on different strands.

The mitochondrial genome exhibits several properties, e.g. simple genetic structure, high rate of evolution, maternal inheritance in most cases [6], making it a useful marker for reconstructing population genetics, phylogenetics and molecular evolution [7, 8]. The nucleotide sequence of single or even multiple genes is frequently used to explore the phylogenetic relationships at the species, genera or family level, while complete mitogenomes have been most useful for phylogenetic analyses above the family level [3, 9-11]. Moreover patterns of mitochondrial gene rearrangements, including gene duplication and random or non-random loss of one of the duplicated elements [6, 12] are well characterized across the metazoan kingdom and were proven useful as apomorphic characters in a phylogenetic context, reviewed in [13].

Within the phylum Arthropoda, complete mitochondrial genomes of more than 270 species are available of which at least 170 hexapod mitochondrial genome sequences are available from the GenBank nonredundant database to date. However, within hexapods there is a strong bias to the derived orders. Although 149 genomes for Neopteran species are available, there is no sequence information for a proturan and a lack of sequence information also at the base of the winged insect still exists. Zhang et al. [14] published the first complete mitochondrial genome of a mayfly which presents the second mt genome of a palaeopteran representative, while the published mt genome of an odonate (*Orthetrum triangulare*) is incomplete [15]. Up to now two additional mt genomes of odonates (*Davidius lunatus* [16] and *Pseudolestes mirabilis* (FJ606784; unpublished)) and two additional mt genomes of mayflies (*Ephemera orientalis* [16], *Siphonurus immanis*

(NC\_013822; unpublished)) are available. In this study, we present two new mitochondrial genomes for representatives of each of the two palaeopterous orders to improve the available taxon sampling of basal pterygote orders: the complete mitochondrial genome of the dragonfly *Boyeria irene* and the nearly complete mitochondrial genome of the mayfly *Baetis* sp. These additional mitochondrial genomes of palaeopterous orders are of crucial importance to clarify the “Palaeoptera problem” by applying a mitogenomic approach. Dragonflies and mayflies represent the most basal extant lineages of Pterygota (winged insects) but determining their relationships with regard to the neopteran clade is difficult due to the ‘explosive radiation’ once flight evolved, see discussion in [17]. There are three possible scenarios explaining the early evolution of winged insects: (i) the “Chiastomyaria” scenario (Odonata, Ephemeroptera + Neoptera) [18, 19], or (ii) the “Metapterygota” scenario (Ephemeroptera, Odonata + Neoptera) [20], or the monophyletic “Palaeoptera” scenario (Odonata + Ephemeroptera, Neoptera), [21, 22]. The recently published multi-gene approach by Simon et al. [23] provided the first critical step towards formulating a robust hypothesis about the evolution of insect flight by claiming the odonates at the base of the pterygotes, supporting the “Chiastomyaria” hypothesis. Also extensive rRNA phylogenies support a basal position of Odonata and are in agreement with the Chiastomyaria hypothesis [24].

However, Zhang et al. [14] applied a mitogenomic approach to resolve the Palaeoptera hypothesis and supported the “Metapterygota” hypothesis. Although the authors performed extensive phylogenetic analyses (MP, BI, ML at the nucleotide as well as at the amino acid level) their data set is characterized by sparse taxon sampling (2 outgroup mitochondrial genomes and 9 pterygote mitochondrial genomes). The used data set could be influenced by the limited taxon sampling since only one odonate and one mayfly is represented. We improve the taxon sampling and add three odonates and three mayflies to the analyses to investigate the influence of taxon sampling for resolving the Palaeoptera problem using a mitogenomic approach. The study represents the largest data set of mitochondrial genomes yet applied to the question of basal pterygote relationships.

## Methods

### *Molecular methods*

#### *Baetis* sp.

Genomic DNA was isolated from an ethanol preserved animal (larvae) according to a modified standard protocol [25]. Based on EST (Expressed Sequence Tags) sequences for *Baetis* sp. (EMBL Nucleotide Sequence Database Accession Nos FN198828-FN203024) [23] and based on the test sequences of the cDNA-library (not rRNA screened) specific primers for 10 mitochondrial protein coding genes and 16S rRNA of *Baetis* sp. were designed. The primer sequences are given in Additional file1. The designed primers were tested in several combinations via PCR and resulted in six successful primer combinations to amplify the entire mitochondrial genome, except some genes (Figure 1); see Additional file 2 for PCR conditions. PCR products were cut with restriction enzymes (AluI, RsaI, HaeIII; Fermentas) and cloned into the pGEM-T vector (Promega) following the procedure for TA-cloning. Sequencing reactions were carried out in both directions with T7 and SP6 standard primers using BigDye Terminator v1.1. Cycle sequencing products were purified with Sephadex™ G-50 Superfine (GE Healthcare) and sequenced using an ABI PRISM™ 310 Genetic Analyzer.

#### *Boyeria irene*

Genomic DNA was extracted from a small amount of thorax tissue of a single ethanol preserved dragonfly following a standard protocol [25]. Universal primers were used to amplify a 427bp region of the mitochondrial *cox2* gene and 392bp region of the 16S rRNA gene. Based on the sequences, specific *Boyeria* primers were designed for long range PCR: combination P3141+P2246 and P3141comp+P2246comp. The complete mitochondrial genome was amplified via long range PCR in two overlapping fragments. Due to failed cloning attempts of the ~10kb fragment the PCR product was cut with restriction enzyme (Alu I), cloned and sequenced. Based on the sequences additional specific primers were designed to amplify the fragment in four overlapping fragments (Figure 2). All primer sequences are given in Additional file 3 and PCR conditions in Additional file 4. The amplified fragments were directly sequenced or sequenced by primer walking at MacroGen (South Korea). The ~6kb fragment was cut with restriction enzyme (AluI, HaeIII, HincII; Fermentas) and cloned into the pGEM-T vector (Promega) following the procedure for TA-cloning or alternatively cloned using the CloneJET™PCR Cloning kit (Fermentas)

according to the manufacturer's instructions. Sequencing reactions were carried out in both directions with T7 and SP6 standard primers and with pJET1.2 forward and reverse standard primers, respectively, using BigDye Terminator v1.1. Cycle sequencing products were purified with Sephadex™ G-50 Superfine (GE Healthcare) and sequenced using an ABI PRISM™ 310 Genetic Analyzer.

Sequences of *Baetis* sp. and *Boyeria irene* were assembled using Lasergene version 5.00 and manually edited. Transfer RNAs were identified by tRNA-scan SE 1.21 [26] and the locations of 13 protein-coding genes and two rRNA (16S and 12S) genes were determined using BLASTX/BLASTN and by comparison with homologous sequences of other insect mtDNA.

### ***Sequence alignments and data sets***

The nucleotide and putative amino acid regions for each of the 13 mitochondrial protein-coding genes for 172 Hexapoda taxa were extracted from NCBI (Additional file 5). The data set comprised 17 primary wingless hexapods, 8 palaeopterous, 29 polyneopterous, 44 paraneopterous and 76 holometabolous insects. Inferred amino acid sequences of all 13 protein coding genes were separately aligned with MAFFT *L-INSI* [27]. In addition the single alignments were back-translated into the corresponding nucleotide sequences. Third codon position were highly saturated as determined using DAMBE v5.2.5 [28] and excluded for the subsequent analyses. Randomly similar sections for all gene alignments (13 nucleotide alignments (1<sup>st</sup>+2<sup>nd</sup> position) and 13 amino acid alignments) were identified using ALIScore [29] with a window size of 6 and a maximal number of pairwise comparisons ( $r=15,051$ ). Each alignment was masked with ALICUT (<http://www.utilities.zfmk.de>) by excluding all randomly similar alignment positions. The masked alignments were concatenated using FASconCAT [30] into a amino acid alignment (3,266 aa) and a nucleotide alignment (6,606 nt).

Due to computational reasons we decided to reduce the original data set and followed for this propose two approaches. (1) The Split decomposition analysis using Neighbor-Net with the uncorrected *p*-distances (Additional file 8) and also preliminary analysis using maximum parsimony in PAUP\* [31] (Additional file 7) pictured the monophyly of most of the orders. Thus we eliminated all but one each of the Lepidoptera, Coleoptera, Heteropterida, Diptera, Neuroptera, Megaloptera and Isoptera. For Blattodea, Caelifera and

Ensifera we included two representatives of each order. (2) In addition, some insect orders were excluded due to their unorthodox position in previous mt genome analyses, e.g. Hymenoptera, Collembola and Diplura. This left 31 taxa, including two apterygote outgroup orders (Archaeognatha and Zygentoma), four odonates, four mayflies, eleven polyneopteran taxa, two paraneopterous taxa and five holometabolus taxa, for which we had one amino acid and one nucleotide dataset.

### ***Phylogenetic analyses***

We employed partitioned Maximum Likelihood (ML) and Bayesian inference (BI) analyses on the 1<sup>st</sup>+2<sup>nd</sup> position of the nucleotide alignment for PCGs (protein-coding genes) to account for different underlying evolutionary models and parameter estimates for individual protein-coding genes [32]. For the amino acid alignment we performed only partitioned ML analyses due to computational reasons.

ML tree search and rapid bootstrapping (1,000 bootstrap replicates) were applied in one step using RAxML Pthreads 7.0.0 [33]. The ML analyses for the amino acid data set were calculated with the PROTMIX [34] substitution model and the MTREV+F model [35] and for the 1<sup>st</sup>+2<sup>nd</sup> position of the nucleotide alignment with the GTRMIX substitution model. The Bayesian analyses (BI) were performed in MRBAYES v3.1.2 [36] on the nucleotide (1<sup>st</sup>+2<sup>nd</sup>) alignment under the model GTR+G+I. Markov Chain Monte Carlo analyses were performed for 4,000,000 generations using four chains and two independent runs. The Bayesian posterior probability (pP) was estimated by sampling the trees every 100 generations after discarding the first 25% (burnin=10,000). Branches with less than 50% bootstrap support and less than 0.80 posterior probability (pP) were collapsed to form polytomies.

## **Results and discussion**

### ***Mitochondrial genome organization and composition***

#### ***Baetis* sp.**

The nearly entire mt genome of *Baetis* sp. was amplified, representing 11,403 bp sequence. 12 protein-coding genes (PCGs) (*nad2* missing), 15 tRNAs and 16S rRNAs (12S rRNA missing) could be identified and are in expected positions and orientations (Figure 1 and Table 1). A variety of approaches were explored to amplify the control region and the

flanking genes, but none of these were successful. Putative fragments spanning the presumed control region and flanking genes were sequenced but turned out to be potential numts (nuclear mitochondrial pseudogenes) [37], see discussion below. All identified PCGs begin with the typical ATN codon (ATG 6 times, ATT 3 times, ATA and ATC each 1 time). The stop codon is truncated for all PCGs (T or TA, respectively), except for *atp6*, *atp8* and *nad4L*. All identified tRNAs could be folded into the typical cloverleaf secondary structure. A total of five short non-coding regions are distributed over the mt genome, with a maximum size of four nucleotides. Also nine gene overlaps were identified with a maximum overlap of seven nucleotides.

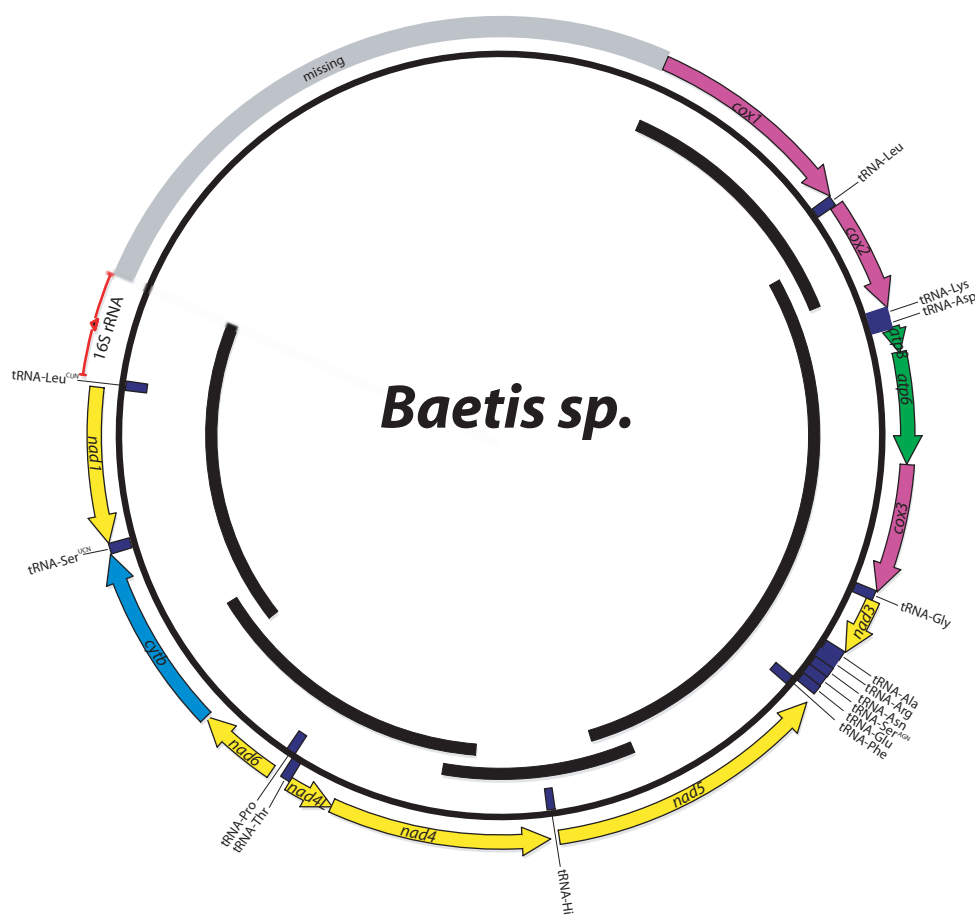
**Table 1**

Annotation of the mitochondrial genome of *Baetis* sp.

Gene or feature	Length (bp)	Start codon	Stop codon	Strand	Intergenic nt
<i>cox1</i>	1279 <sup>a</sup>	missing	T	+	0
tRNA-Leu <sup>UUR</sup>	64			+	4
<i>cox2</i>	680	ATG	TA	+	0
tRNA-Lys	64			+	-1
tRNA-Asp	61			+	0
<i>atp8</i>	162	ATC	TAA	+	-7
<i>atp6</i>	675	ATG	TAA	+	2
<i>cox3</i>	784	ATG	TA	+	-4
tRNA-Gly	62			+	0
<i>nad3</i>	349	ATT	T	+	0
tRNA-Ala	63			+	-1
tRNA-Arg	61			+	-1
tRNA-Asn	64			+	-2
tRNA-Ser <sup>AGN</sup>	68			+	0
tRNA-Glu	61			+	-1
tRNA-Phe	63			-	4
<i>nad5</i>	1725	ATA	TA	-	0
tRNA-His	61			-	-1
<i>nad4</i>	1334	ATG	TA	-	-7
<i>nad4L</i>	297	ATG	TAA	-	1
tRNA-Thr	62			+	0
tRNA-Pro	62			-	1
<i>nad6</i>	509	ATT	TA	+	0
<i>cytb</i>	1136	ATG	TA	+	0
tRNA-Ser <sup>UCN</sup>	65			+	0
<i>nad1</i>	945	ATT	T	-	0
tRNA-Leu <sup>CUN</sup>	61			-	0
16S rRNA	599 <sup>a</sup>			-	

<sup>a</sup> incomplete sequence

\* 7 tRNAs, *nad2* and control region are missing



**Figure 1**

Circular map of the mitochondrial genome of *Baetis* sp, illustrated with Seqbuilder 7.2.2. Transcriptional direction of each protein-coding gene is indicated by the direction of the arrow. tRNAs are denoted as three-letter symbol according to the IUPAC-IUB amino acid code. tRNAs encoded by the L-strand are shown in the inner circle and tRNAs of the H-strand in the outer circle. The black overlapping lines in the inner circle illustrate the amplified fragments. The control region, 12S rRNA, *nad2* and 7 tRNAs are missing. 16S rRNA and *cox1* are incomplete.

### *Boyeria irene*

Although the entire mt genome of *Boyeria irene* was amplified, we failed in determining the control region and flanking genes due to non overlapping subclones of the restriction enzyme digestion. The control region spanning fragment was successfully amplified after a second approach with different primer combinations. The first approach resulted into the amplification of a potential numt (see discussion below). The until now determined *B. irene* mt genome consist of 13,638 bp, harbouring the 13 PCGs, 18 tRNAs and 2 rRNAs (16S rRNA and 12S rRNA) with the same orientation and gene order common to insects [2] (Figure 2 and Table 2). The identified PCGs begin with the typical ATN codon (ATG 5 times, ATT 3 times, ATA and ATC each 1 time), except for *nad1* which starts with TTG. Although 27 nucleotides downstream of the predicted ORF a possible ATA start codon is



located, the TTG start codon appears more plausible in an evolutionary economic way, minimizing the intergenic space between *cox1* and tRNA-Ser [38]. In addition, TTG has also been proposed as start codon for *cox1* in several insects [38-41]. Nine non-coding regions were identified with a maximum size of 22 nucleotides and six gene overlaps with a maximum size of four nucleotides.

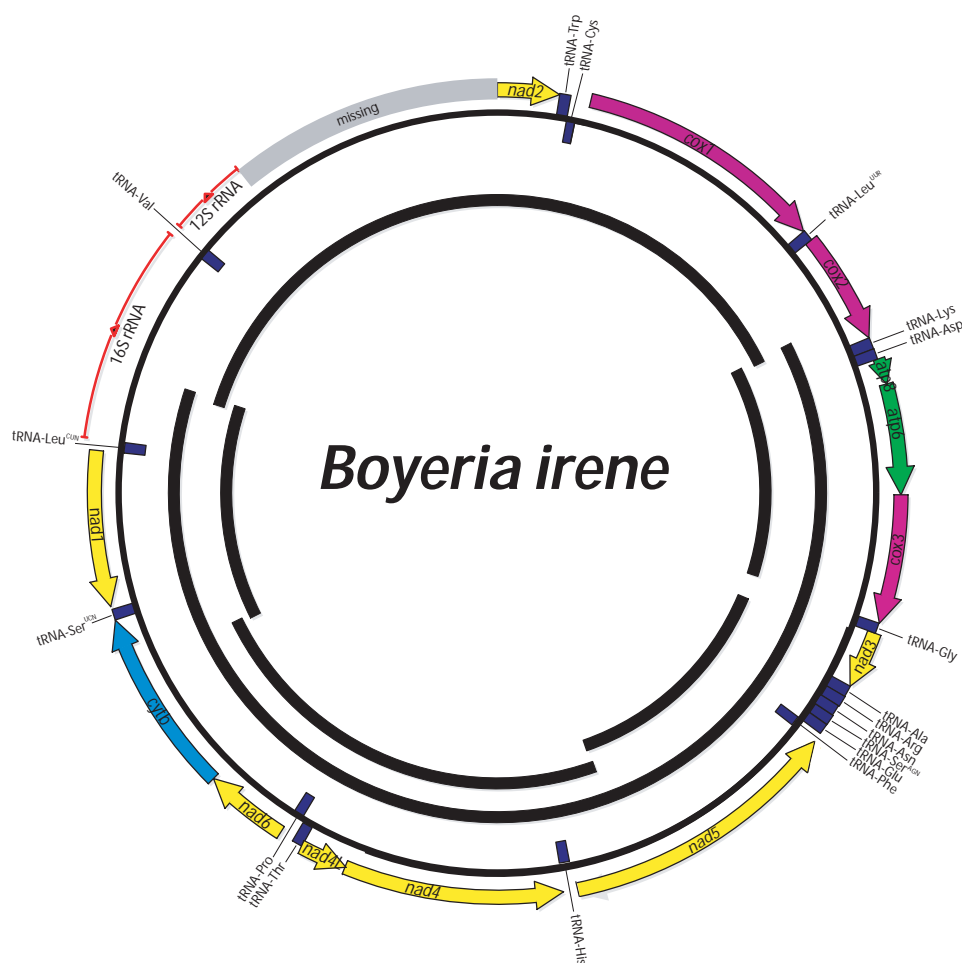
**Table 2**

Annotation of the mitochondrial genome of *Boyeria irene*.

Gene or feature	Length (bp)	Start codon	Stop codon	Strand	Intergenic nt
<i>nad2</i>	379 <sup>a</sup>	missing	T	+	0
tRNA-Trp	69			+	0
tRNA-Cys	55			-	0
tRNA-Tyr	missing				-
<i>cox1</i>	1537	missing	T	+	0
tRNA-Leu <sup>UUR</sup>	69			+	0
<i>cox2</i>	688	ATG	T	+	0
tRNA-Lys	72			+	0
tRNA-Asp	66			+	1
<i>atp8</i>	162	ATT	TAA	+	-4
<i>atp6</i>	675	ATA	TAA	+	-1
<i>cox3</i>	787	ATG	TA	+	-1
tRNA-Gly	65			+	0
<i>nad3</i>	354	ATT	TAA	+	-1
tRNA-Ala	71			+	-1
tRNA-Arg	68			+	0
tRNA-Asn	66			+	1
tRNA-Ser <sup>AGN</sup>	66			+	1
tRNA-Glu	66			+	0
tRNA-Phe	65			-	0
<i>nad5</i>	1730	ATT	TA	-	0
tRNA-His	67			-	3
<i>nad4</i>	1344	ATG	TAA	-	-1
<i>nad4L</i>	289	ATG	T	-	2
tRNA-Thr	66			+	22
tRNA-Pro	66			-	UUR
<i>nad6</i>	521	ATC	TA	+	0
<i>cytb</i>	1134	ATG	TAA	+	1
tRNA-Ser <sup>UCN</sup>	67			+	18
<i>nad1</i>	951	TTG	TAA	-	1
tRNA-Leu <sup>CUN</sup>	68			-	0
16S rRNA	1300			-	0
tRNA-Val	71			-	0
12S rRNA	476 <sup>a</sup>			-	

<sup>a</sup> incomplete sequence

\* 4 tRNAs and control region are missing



### Figure 2

Circular map of the mitochondrial genome of *Boyeria irene*, illustrated with Seqbuilder 7.2.2. Transcriptional direction of each protein-coding gene is indicated by the direction of the arrow. tRNAs are denoted as three-letter symbol according to the IUPAC-IUB amino acid code. tRNAs encoded by the L-strand are shown in the inner circle and tRNAs of the H-strand in the outer circle. The black overlapping lines in the inner circle illustrate the amplified fragments. The control region and 4 tRNAs are missing. 12S rRNA, *cox1* and *nad2* are incomplete.

## Numts

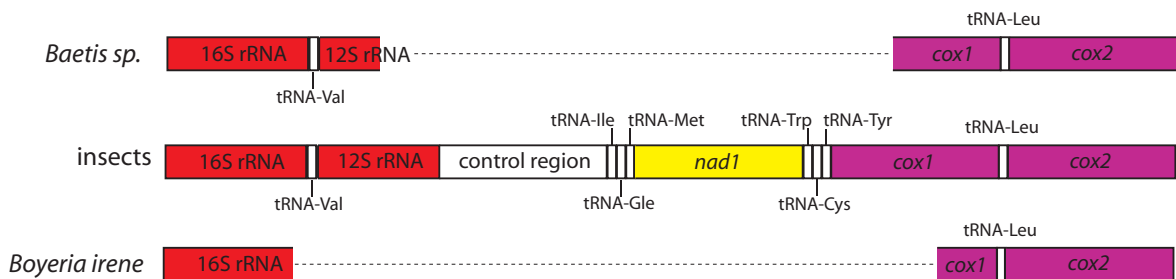
The phenomenon of gene transfer from the mitochondrial genome into the nuclear genome has been reported for several insects, e.g. for grasshoppers, aphids, flies, beetles, bees and mosquitoes [42-47], but they have not been described so far for palaeopterous insects. We discovered in both palaeopterous insect orders a numt, potentially derived from the control region and flanking genes with subsequent deletions (Figure 3). The loss of function of the numts was observed confirmed by a large deletion in functionally important coding sequences.

Through BLASTX/BLASTN and comparison to other published *cox1* sequences, it turned out that in the numt of *Baetis* sp. the *cox1* gene is truncated, followed directly by 12S rRNA. This could indicate that a larger fragment have been transferred first into the

nuclear genome with a subsequent deletion. It has been shown that deletions occur frequently in numts contributing greatly to the degradation of these pseudogenes [45]. We observed a similar numt in *Boyeria irene*. The amplified fragment (pseudogene) between the 16S rRNA and the *cox2* gene had a size of approx. 1.3 kb. Within the sequence, we could identify *cox2*, the tRNA-Leu, truncated *cox1*, and 16S rRNA. However, the sequences of the control region, the genes *nad1*, 12S rRNA and seven tRNAs (valine, isoleucine, glycine, methionine, tryptophan, cysteine and tyrosine), typically located between 16S rRNA and *cox2*, were missing (Figure 3).

The discovery of numts in both palaeopterous basal insect orders indicates that they are even more widespread in insects as previously suggested [45]. Moreover, our results indicate that nuclear locations for mitochondrial-like sequences of the fragment *cox1/cox2*, *nad2*, the rRNAs and intermediate tRNAs might have a numerous distribution in insects, as already suggested by [42, 48].

We can not unambiguously rule out that the two observed numts are the only ones which can be amplified in both taxa. However, we have reduced the chance to amplify numts by performing long range PCRs [49] and the amplified fragments with correct open reading frames (ORF's) correspond to the expected genes. Consequently, we expect the amplification of mitochondrial genes. But still, the discovery of numts in several insect orders requires more attention and care in conducting phylogenetic analyses based on amplified mitochondrial genes. The inclusion of numt sequences can lead to erroneous phylogenetic inferences and obscures the evolutionary history of the species [47].



**Figure 3**

Illustration of amplified numts in *Baetis* sp. and *Boyeria irene*. Shown is the common gene arrangement of the region between 16S rRNA and *cox2* for insect mitochondrial genomes in comparison to the amplified numts in *Baetis* sp. and *Boyeria irene*. Above: identified numt in *Baetis* sp.; dashed line indicate missing part in the amplified nuclear pseudogene. Below: identified numt in *Boyeria irene*; dashed line indicate missing part in the amplified nuclear pseudogene.

### ***Alignment masking***

The high evolutionary rate of mitochondrial genomes can cause saturation of the phylogenetic signal, especially problematic in deep split phylogenies. Moreover, highly divergent sequences are often only ambiguously alignable and contain random similarity due to the mutational saturation [29, 50]. These random similar positions in the alignment can bias tree reconstructions in several ways and exclusion of these saturated regions can help to reduce noise [29, 51, 52]. In order to identify saturated regions within the gene alignments ALISCORE [29] was applied. Within the amino acid alignment the fewest random similar positions were identified in *cytb* and *cox1-cox3*. Although *atp8* is usually a priori excluded from the analyses in most studies due to sequence heterogeneity [11, 14], less than 50% were identified as randomly similar positions. In contrast, for *nad6* the highest percent of randomly similar positions were identified (Additional file 6). Based on the 1<sup>st</sup>+2<sup>nd</sup> nucleotide gene alignments, ALISCORE identified less than 40% randomly similar positions in *atp8* and the highest percent of randomly similar positions in *nad2*.

### **Phylogenetic relationships**

The substantial availability of mitochondrial genome sequences within the superclass Hexapoda provides an important data source for deep level phylogenies. However, no mitogenomic approach was conducted on the complete data set; instead several studies had limited scope inferring intraordinal relationships or interordinal relationships with the infraclasses (Polyneoptera, Paraneoptera and Holometabola). Cameron et al. [53] conducted a mitogenomic approach inferring relationships within Neuropterida and between Neuropterida and other holometabolous insect orders but excluded several hymenopteran species from the analyses due to compositional bias. Also Castro and Downton [54] excluded several insect mitochondrial genome sequences prior to the analyses because of compositional bias and long-branch attraction. In contrast, mitogenomic approaches using a broader taxon sampling across hexapods have produced some unexpected results. Nardi et al. [9] and Cook et al. [11], e.g. supported a polyphyly of Hexapoda with collembolans emerging before crustaceans and evolving separately from other insects. Carapelli et al. [55] confirmed the paraphyly of hexapods and supported further a paraphyly of Endopterygota with Diptera as a sister group to Plecoptera based on a mitogenomic approach. These results contradict both morphological and molecular studies in hexapods and indicate the potential pitfalls in the analysis of mt genome data. (1) Mitochondrial DNA evolves more rapidly than nuclear DNA causing mutational saturation

which becomes a major problem for deeper nodes [56]. (2) The compositional heterogeneity among lineages and genes is known to bias the performance of phylogenetic inference methods [57]. (3) The heterogeneity of evolutionary rates among sites can introduce long-branch attraction (LBA) artefacts [58]. Cameron et al. [59] reviewed these existing problems in mitogenomic phylogenies inferring deep intraordinal relationships among insects. However, the authors did not present the inferred insect relationships based on mitochondrial genomes comprising 29 insect orders.

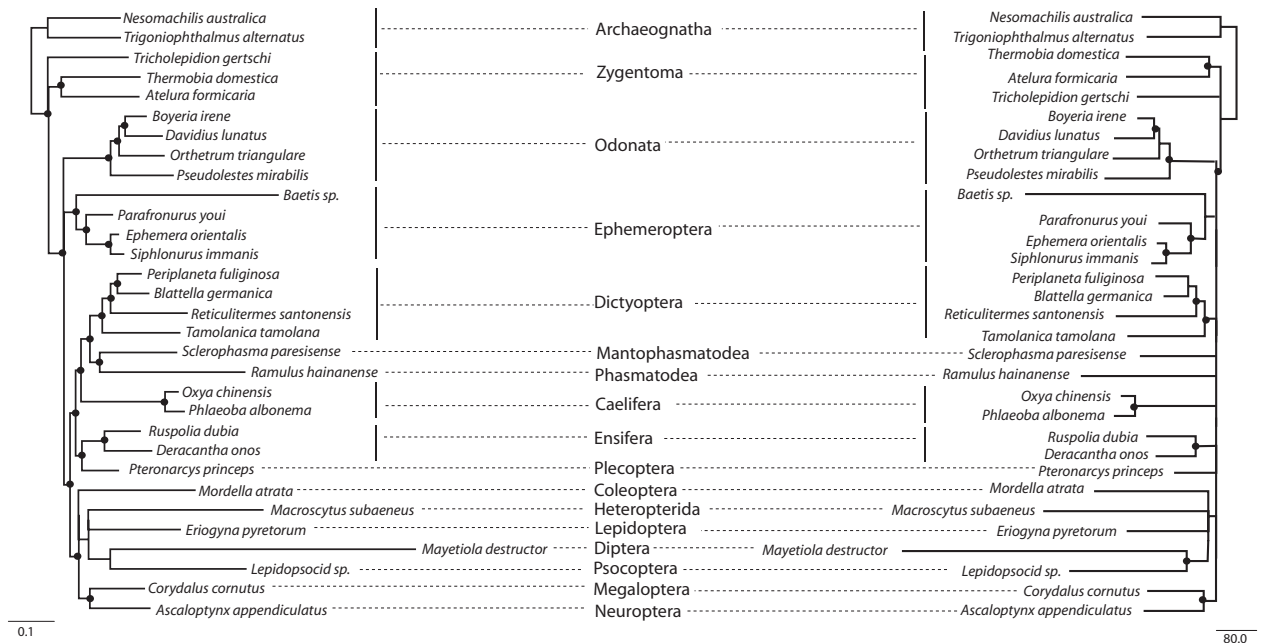
Due to computational reasons, we performed only Maximum Parsimony (MP) analysis on the amino acid alignment of the original data set (174 hexapod species). The inferred topology based on 1000 bootstrap replicates using PAUP\* [31] demonstrates the overall lack of phylogenetic signal for the deep clades among hexapods (Additional file 7). High support was only recovered for the monophyly of some hexapod orders: Entognatha (0.95), Archaeognatha (0.98), Odonata (1.0), Caelifera (1.0) and Lepidoptera (1.0). The lack of phylogenetic signal for the deep nodes is also reflected in the Neighbor-Net reconstruction. The Neighbor-Net graph which results from a split decomposition based on uncorrected *p*-distances is non-treelike, and shows short and conflicting internal branches and long terminal branches (Additional file 8).

However, the purpose of this study was to assess the stability of the Metapterygota hypothesis supported by a previous mitogenomic approach of Zhang et al. [14]. In contrast, phylogenetic studies based on nuclear genes, rRNAs and multi-gene analyses support the Chistomyaria hypothesis, e.g. [23, 24, 60, 61]. Here, we improved the taxon sampling fourfold for both palaeopterous orders (Ephemeroptera and Odonata) and added several ingroup and outgroup taxa for the mitogenomic approach.

In the analyses of mitochondrial nucleotide and amino acid data sets a number of insect relationships is not stable. Unfortunately, these unstable positions among the inferred topologies are also related to the palaeopterous orders, consequently leaving the Palaeoptera problem – the primary interest of this study – unresolved.

The BI and ML analyses based on the 1<sup>st</sup>+2<sup>nd</sup> position of the PCGs generated overall similar topologies, except for the position of *Tricholepidion gertschi* (Figure 4A and B). However, the bootstrap support values were below 50% and consequently collapsed. Although monophyletic Pterygota were well recovered in both analyses (pP: 1.0 and

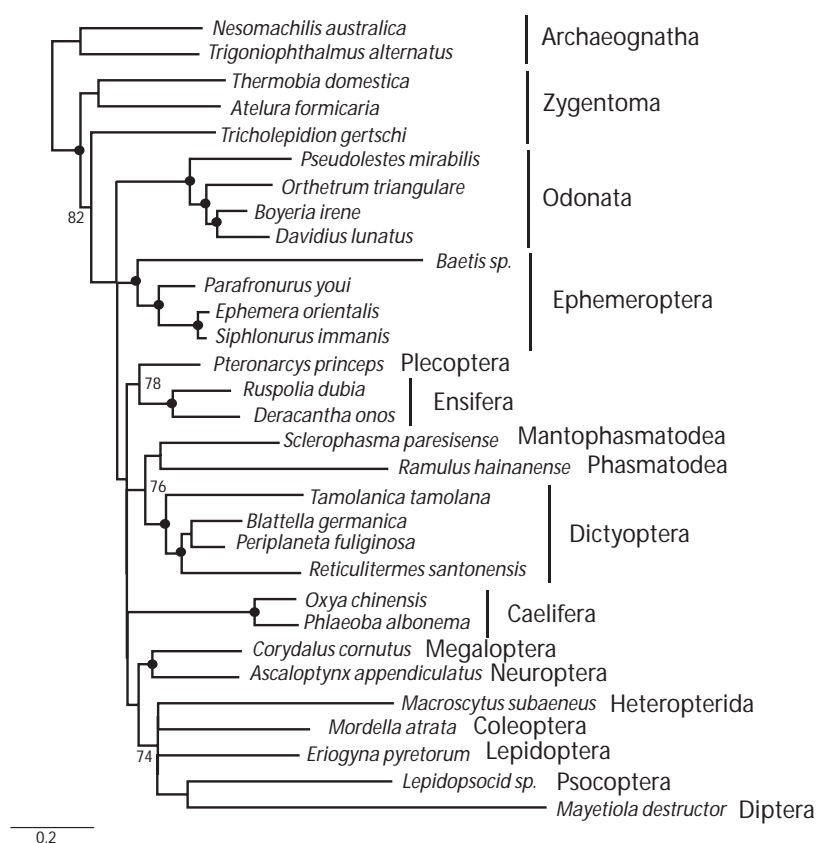
bootstrap support: 100), there was no significantly support for the earliest branching lineage of Pterygota (bootstrap support <50% and pP <0.90). A monophyletic Neoptera (pP=1.0) and Polyneoptera (pP=1.0) were only recovered in the BI analyses. The two inferred topologies well supported a monophyly of Odonata, Ephemeroptera, Blattodea, Caelifera and Ensifera. A monophyletic Dictyoptera (Mantodea+Isoptera+Blattodea) was well recovered in both ML and BI analyses. A sister-group relationship of Mantophasmatodea and Phasmatodea was only supported in the BI analysis (pP 0.99), congruent to the results of [62]. On the other hand, the relationships between the holometabolous and paraneopterous orders were weakly supported in both analyses. The well supported monophyly of Holometabola was disrupted by the paraneopterous taxa, also seen in [22, 63, 64]. Moreover, we could not recover a monophyletic Orthoptera (Ensifera+Caelifera), as observed in other mitogenomic analyses [55, 65]. In this context, Ma et al. [66] demonstrated that mitogenomic approaches fail to recover a monophyletic Orthoptera if only limited orthopteran species are represented.



**Figure 4**

Phylogenetic trees obtained from mitochondrial nucleotide data set; 1<sup>st</sup>+2<sup>nd</sup> position of PCGs and exclusion of randomly identified positions. Left: BI tree; pP < 0.90: not shown; pP ≥ 0.99: represented by a dot only. Right: ML tree (majority rule consensus). Support values < 70: not shown, support values >90: represented by a dot only.

The inferred topology of the partitioned ML analyses based on the amino acid data set showed overall low support for the deep splits. Monophyletic Pterygota was well supported (bootstrap support=100%), but again there was no significant support for the earliest branching lineage of Pterygota. In addition we could not recover monophyletic Neoptera or Polyneoptera. The following intraordinal relationships were well supported: monophyly of Ephemeroptera, Odonata, Ensifera, Caelifera and Blattodea. High support was also recovered for the Dictyoptera clade (Mantodea+Isoptera+Blattodea). In the inferred topology based on the amino acid data set, a monophyly of Orthoptera (Ensifera+Caelifera) was not supported and the monophyly of Holometabola was disrupted by paraneopterous taxa (Figure 5).



**Figure 5**

ML tree obtained from mitochondrial amino acid data set. Support values < 70: not shown, support values >90: represented by a dot only.

Until now, we can not assess the lack of support for the deep nodes especially in the ML analyses for both data sets. The analyses showed that the inference method had a significant effect on the phylogenetic reconstruction. The BI analyses were more consistent with previous inferred phylogenetic relationships based on other molecular markers. In

addition, the analyses based on the nucleotide alignment resulted to more reasonable inferred insect relationships than the analyses based on the amino acid alignment. Perhaps we will obtain a better resolution by combining all mt genome data, PCGs, tRNAs and rRNAs. Studies have demonstrated that the best way for analyzing mitogenomic data is to perform the phylogenetic analyses at the nucleotide level and include all available data [67].

However, it remains to be seen if the Palaeoptera problem could be resolved robustly by a mitogenomic approach. Most studies have shown that mitogenomic data performs well targeting divergence times from the deep Permian (~260 MYA) to the Tertiary (~50 MYA) [67], but the usefulness of mitogenomic data for inferring relationships for highly divergent lineages is still controversial. Lin and Danforth [68] argue that nuclear genes data sets should be preferred for deep insect molecular phylogenetics due to the substitutional biases and high evolutionary rate of mitochondrial genes. In this context, Timmermans et al. [69] highlighted the discrepancies between inferred relationships of nuclear and mitochondrial data and recovered the well supported monophyly of Hexapoda with nuclear data. Instead, mitogenomic approaches still failed to recover monophyletic Hexapoda [40, 55].

However, other studies have shown that mitochondrial genomes could be useful for deep phylogenetic relationships and are able to retrieve plausible phylogenetic relationships by applying appropriate substitution models [70, 71]. Still, these substitution models for mitochondrial sequence evolution might be first developed to robustly resolve the earliest branching lineages of Pterygota using mitogenomic data.

## **Conclusions**

Although a previous mitogenomic approach confirmed the Metapterygota hypothesis, the phylogenetic inference with the evidence from additional mitochondrial genomes in this study indicates the sensitivity of taxon sampling to the inferred relationships. The present study demonstrates the high impact of additional mitochondrial genome sequences to inferred deep level insect relationships. Furthermore, our analyses show that from a mitogenomic point of view the highly debated phylogeny among basal winged insect remains unresolved so far.



**Abbreviations**

*atp6* and *atp8*: ATP synthase FO subunit 6 and 8 genes; *cox1-3*: cytochrome c oxidase subunit I, II and III genes; *cytb*: cytochrome b gene; *nad1-6 nad4L*: NADH dehydrogenase subunit 1-6 and 4L genes; tRNA: transfer RNA. PCG: protein-coding genes; aa: amino acid; nt: nucleotide; MP: Maximum Parsimony; ML: Maximum Likelihood; BI: Bayesian Inference; MYA: million years ago.

**Author contributions**

SS was primarily responsible for the design of the study, conducted the experiments and analyses and drafted the primary version of the manuscript. Both authors discussed and approved the final manuscript.

**Additional material****Additional file1**

Primer list for *Baetis* sp.

**Additional file2**

PCR conditions for *Baetis* sp.

**Additional file 3**

Primer list for *Boyeria irene*.

**Additional file 4**

PCR conditions for *Boyeria irene*.

**Additional file 5**

Taxa list. Taxa list of mitochondrial genomes of hexapods. \* indicates taxa included in reduced data set.

**Additional file 6**

Randomly similar identified positions in the single mitochondrial gene alignments. Given are the percentages left after masking the single gene alignments (1<sup>st</sup>+2<sup>nd</sup> nucleotide positions of PCGs and amino acids) with ALICUT (<http://www.utilities.zfmk.de>) by excluding all randomly similar alignment positions identified by ALIScore.

**Additional file 7**

MP topology inferred from the amino acid data set with all 174 mt genomes.

**Additional file 8**

Neighbor-Net graph based on split decomposition with the uncorrected  $p$ -distance of the amino acid alignment of all 174 hexapods mtgenomes using SplitsTree4 after exclusion of randomly similar sections evaluated with ALISCORE.

## Acknowledgements

We thank Jennifer Angermann and Ann Kathrin Ketelsen for their support with the laboratory work. This work was supported by the German Science Foundation (DFG) in the priority program SPP 1174 "Deep Metazoan Phylogeny", DFG grant HA 1947/5.

## References

1. Lavrov DV: **Key transitions in animal evolution: a mitochondrial DNA perspective.** *Integrative and Comparative Biology* 2007, **47**(5):734-743.
2. Boore J: **Animal mitochondrial genomes.** *Nucleic Acids Res* 1999, **27**:1767-1780.
3. Podsiadlowski L, Carapelli A, Nardi F, Dallai R, Koch M, Boore JL, Frati F: **The mitochondrial genomes of *Campodea fragilis* and *Campodea lubbocki* (Hexapoda: Diplura): High genetic divergence in a morphologically uniform taxon.** *Gene* 2006, **381**:49-61.
4. Saito S, Tamura K, Aotsuka T: **Replication origin of mitochondrial DNA in insects.** *Genetics* 2005, **171**(4):1695-1705.
5. Bender A, Hajieva P, Moosmann B: **Adaptive antioxidant methionine accumulation in respiratory chain complexes explains the use of a deviant genetic code in mitochondria.** *P Natl Acad Sci USA* 2008, **105**(43):16496-16501.
6. Moritz C, Dowling TE, Brown WM: **Evolution of Animal Mitochondrial-DNA - Relevance for Population Biology and Systematics.** *Annu Rev Ecol Syst* 1987, **18**:269-292.
7. Kim KG, Hong MY, Kim MJ, Im HH, Kim MI, Bae CH, Seo SJ, Lee SH, Kim I: **Complete mitochondrial genome sequence of the yellow-spotted long-horned beetle *Psacotha hilaris* (Coleoptera: Cerambycidae) and phylogenetic analysis among coleopteran insects.** *Mol Cells* 2009, **27**(4):429-441.
8. Lee W, Kang J, Jung C, Hoelmer K, Lee SH, Lee S: **Complete mitochondrial genome of brown marmorated stink bug *Halyomorpha halys* (Hemiptera: Pentatomidae), and phylogenetic relationships of hemipteran suborders.** *Mol Cells* 2009.
9. Nardi F, Spinsanti G, Boore JL, Carapelli A, Dallai R, Frati F: **Hexapod origins: monophyletic or paraphyletic?** *Science* 2003, **299**(5614):1887-1889.
10. Cameron SL, Miller KB, D'Haese CA, Whiting MF, Barker SC: **Mitochondrial genome data alone are not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea \textit{sensu lato} (Arthropoda).** *Cladistics* 2004, **20**(6):534-557.
11. Cook CE, Yue Q, Akam M: **Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic.** *Proc Biol Sci* 2005, **272**(1569):1295-1304.
12. Lavrov DV, Boore JL, Brown WM: **Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: duplication and nonrandom loss.** *Mol Biol Evol* 2002, **19**(2):163-169.

13. Boore JL: **The use of genome-level characters for phylogenetic reconstruction.** *Trends Ecol Evol* 2006, **21**(8):439-446.
14. Zhang J, Zhou C, Gai Y, Song D, Zhou K: **The complete mitochondrial genome of *Parafronurus youi* (Insecta: Ephemeroptera) and phylogenetic position of the Ephemeroptera.** *Gene* 2008, **424**(1-2):18-24.
15. Yamauchi MM, Miya MU, Nishida M: **Use of a PCR-based approach for sequencing whole mitochondrial genomes of insects: two examples (cockroach and dragonfly) based on the method developed for decapod crustaceans.** *Insect Mol Biol* 2004, **13**(4):435-442.
16. Lee EM, Hong MY, Kim MI, Kim MJ, Park HC, Kim KY, Lee IH, Bae CH, Jin BR, Kim I: **The complete mitogenome sequences of the palaeopteran insects *Ephemera orientalis* (Ephemeroptera: Ephemeridae) and *Davidius lunatus* (Odonata: Gomphidae).** *Genome* 2009, **52**(9):810-817.
17. Whitfield JB, Kjer KM: **Ancient rapid radiations of insects: challenges for phylogenetic analysis.** *Annu Rev Entomol* 2008, **53**:449-472.
18. Boudreaux BH: **Arthropod phylogeny: with special reference to insects.** John Wiley & Sons Inc.; 1979.
19. Kjer KM: **Aligned 18S and insect phylogeny.** *Syst Biol* 2004, **53**(3):506-514.
20. Kristensen NP: **Phylogeny of extant hexapods.** In: *The Insects of Australia: A Textbook for Students and Research Workers.* Edited by Naumann ID, Lawrence, J.F., Nielsen, E.S., Spradberry, J.P., Taylor, R.W., Whitten, M.J., Littlejohn, M.J., vol. 125-140. Melbourne: CSIRO, Melbourne Univ. Press; 1991.
21. Hennig W: **Insect Phylogeny.** New York: John Wiley & Sons; 1981.
22. Kukalova-Peck J: **Fossil history and the evolution of hexapod structures.** In: *The Insects of Australia: A Textbook for Students and Researcher Workers.* Edited by Naumann ID, Lawrence, J.F., Nielsen, E.S., Spradberry, J.P., Taylor, R.W., Whitten, M.J., Littlejohn, M.J. Melbourne: CSIRO Melbourne University Press; 1991: 141-179.
23. Simon S, Strauss S, von Haeseler A, Hadrys H: **A phylogenomic approach to resolve the basal pterygote divergence.** *Mol Biol Evol* 2009, **26**(12):2719-2730.
24. von Reumont BM, Meusemann K, Szucsich NU, Dell'Ampio E, Gowri-Shankar V, Bartel D, Simon S, Letsch HO, Stocsits RR, Luan YX *et al*: **Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships.** *BMC Evol Biol* 2009, **9**:119.
25. Hadrys H, Schierwater B, Dellaporta SL, DeSalle R, Buss LW: **Determination of paternity in dragonflies by Random Amplified Polymorphic DNA fingerprinting.** *Mol Ecol* 1993, **2**(2):79-87.
26. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25**(5):955-964.
27. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program.** *Brief Bioinform* 2008, **9**(4):286-298.
28. Xia X, Xie Z: **DAMBE: software package for data analysis in molecular biology and evolution.** *J Hered* 2001, **92**(4):371-373.
29. Misof B, Misof K: **A Monte Carlo Approach Successfully Identifies Randomness in Multiple Sequence Alignments : A More Objective Means of Data Exclusion.** *Systematic Biol* 2009, **58**(1):21-34.
30. Kück P, Meusemann K: **FASconCAT: Convenient handling of data matrices.** *Mol Phylogenet Evol.*

31. Swofford DL: **PAUP\* Phylogenetic Analysis Using Parsimony (\*and Other Methods)**. In, 4 edn. Sunderland, MA: Sinauer Associates; 2002.
32. Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JeL: **Bayesian phylogenetic analysis of combined data**. *Syst Biol* 2004, **53**(21):47-67.
33. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics* 2006, **22**(21):2688-2690.
34. Stamatakis A: **Phylogenetic models of rate heterogeneity: A high performance computing perspective**. In: *20th International Parallel and Distributed Processing Symposium (IPDPS 2006), Proceedings: 25-29 April 2006; Rhodes Island, Greece; 2006*.
35. Adachi J, Hasegawa M: **Model of amino acid substitution in proteins encoded by mitochondrial DNA**. *J Mol Evol* 1996, **42**(4):459-468.
36. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models**. *Bioinformatics* 2003, **19**(12):1572-1574.
37. Song H, Buhay JE, Whiting MF, Crandall KA: **Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified**. *Proc Natl Acad Sci U S A* 2008, **105**(36):13486-13491.
38. Bae JS, Kim I, Sohn HD, Jin BR: **The mitochondrial genome of the firefly, *Pyrocoelia rufa*: complete DNA sequence, genome organization, and phylogenetic analysis with other insects**. *Mol Phylogenet Evol* 2004, **32**(3):978-985.
39. Mitchell SE, Cockburn AF, Seawright JA: **The mitochondrial genome of *Anopheles quadrimaculatus* species A: complete nucleotide sequence and gene organization**. *Genome* 1993, **36**(6):1058-1073.
40. Nardi F, Carapelli A, Dallai R, Frati F: **The mitochondrial genome of the olive fly *Bactrocera oleae*: two haplotypes from distant geographical locations**. *Insect Mol Biol* 2003, **12**(6):605-611.
41. Sheffield NC, Song H, Cameron SL, Whiting MF: **A comparative analysis of mitochondrial genomes in Coleoptera (Arthropoda: Insecta) and genome descriptions of six new beetles**. *Mol Biol Evol* 2008, **25**(11):2499-2509.
42. Bensasson D, Zhang DX, Hewitt GM: **Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes**. *Mol Biol Evol* 2000, **17**(3):406-415.
43. Sunnucks P, Hales DF: **Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus *Sitobion* (Hemiptera: Aphididae)**. *Mol Biol Evol* 1996, **13**(3):510-524.
44. Pamilo P, Viljakainen L, Vihavainen A: **Exceptionally high density of NUMTs in the honeybee genome**. *Mol Biol Evol* 2007, **24**(6):1340-1346.
45. Pons J, Vogler AP: **Complex pattern of coalescence and fast evolution of a mitochondrial rRNA pseudogene in a recent radiation of tiger beetles**. *Mol Biol Evol* 2005, **22**(4):991-1000.
46. Richly E, Leister D: **NUMTs in sequenced eukaryotic genomes**. *Mol Biol Evol* 2004, **21**(6):1081-1084.
47. Hlaing T, Tun-Lin W, Somboon P, Socheat D, Setha T, Min S, Chang MS, Walton C: **Mitochondrial pseudogenes in the nuclear genome of *Aedes aegypti* mosquitoes: implications for past and future population genetic studies**. *BMC Genet* 2009, **10**:11.
48. Vaughan HE, Heslop-Harrison JS, Hewitt GM: **The localization of mitochondrial sequences to chromosomal DNA in orthopterans**. *Genome* 1999, **42**(5):874-880.

49. Akanuma J, Muraki K, Komaki H, Nonaka I, Goto Y: **Two pathogenic point mutations exist in the authentic mitochondrial genome, not in the nuclear pseudogene.** *J Hum Genet* 2000, **45**(6):337-341.
50. Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Wagele JW, Misof B: **Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees.** *Front Zool* 2010, **7**:10.
51. Wägele JW, Mayer C: **Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects.** *BMC Evol Biol* 2007, **7**(1):147.
52. Dress AW, Flamm C, Fritzsche G, Grunewald S, Kruspe M, Prohaska SJ, Stadler PF: **Noisy: identification of problematic columns in multiple sequence alignments.** *Algorithms Mol Biol* 2008, **3**:7.
53. Cameron SL, Sullivan J, Song HJ, Miller KB, Whiting MF: **A mitochondrial genome phylogeny of the Neuropterida (lace-wings, alderflies and snakeflies) and their relationship to the other holometabolous insect orders.** *Zoologica Scripta* 2009, **38**(6):575-590.
54. Castro LR, Dowton M: **The position of the Hymenoptera within the Holometabola as inferred from the mitochondrial genome of *Perga condei* (Hymenoptera: Symphyta: Pergidae).** *Mol Phylogenet Evol* 2005, **34**(3):469-479.
55. Carapelli A, Li'o P, Nardi F, van der Wath E, Frati F: **Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea.** *BMC Evol Biol* 2007, **Suppl 2**:S8.
56. Graur D, Li WH: **Fundamentals of molecular evolution.** Sunderland, Massachusetts: Sinauer Associates; 2000.
57. Gibson A, Gowri-Shankar V, Higgs PG, Rattray M: **A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods.** *Molecular Biology and Evolution* 2005, **22**(2):251-264.
58. Reyes A, Pesole G, Saccone C: **Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny.** *Gene* 2000, **259**(1-2):177-187.
59. Cameron SL, Beckenbach AT, Dowton M, Whiting MF: **Evidence from Mitochondrial Genomics on Interordinal Relationships in Insects.** *Arthropod Systematics & Phylogeny* 2006, **64**(1):27-34.
60. Misof B, Niehuis O, Bischoff I, Rickert A, Erpenbeck D, Staniczek A: **Towards an 18S phylogeny of hexapods: accounting for group-specific character covariance in optimized mixed nucleotide/doublet models.** *Zoology (Jena)* 2007, **110**(5):409-429.
61. Kjer KM: **Aligned 18S and insect phylogeny.** *Syst Biol* 2004, **53**(3):506-514.
62. Cameron SL, Barker SC, Whiting MF: **Mitochondrial genomics and the new insect order Mantophasmatodea.** *Mol Phylogenet Evol* 2006, **38**(1):274-279.
63. Whiting MF, Carpenter JC, Wheeler QD, Wheeler WC: **The Strepsiptera problem: phylogeny of the holometabolous insect orders inferred from 18S and 28S ribosomal DNA sequences and morphology.** *Syst Biol* 1997, **46**(1):1-68.
64. Stewart JB, Beckenbach AT: **Phylogenetic and genomic analysis of the complete mitochondrial DNA sequence of the spotted asparagus beetle *Crioceris duodecimpunctata*.** *Mol Phylogenet Evol* 2003, **26**(3):513-526.
65. Kim I, Cha SY, Yoon MH, Hwang JS, Lee SM, Sohn HD, Jin BR: **The complete nucleotide sequence and gene organization of the mitochondrial genome of the oriental mole cricket, *Gryllotalpa orientalis* (Orthoptera: Gryllotalpidae).** *Gene* 2005, **353**(2):155-168.

66. Ma C, Liu C, Yang P, Kang L: **The complete mitochondrial genomes of two band-winged grasshoppers, *Gastrimargus marmoratus* and *Oedaleus asiaticus*.** *BMC Genomics* 2009, **10**:156.
67. Fenn JD, Song H, Cameron SL, Whiting MF: **A preliminary mitochondrial genome phylogeny of Orthoptera (Insecta) and approaches to maximizing phylogenetic signal found within mitochondrial genome data.** *Mol Phylogenet Evol* 2008, **49**(1):59-68.
68. Lin CP, Danforth BN: **How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets.** *Mol Phylogenet Evol* 2004, **30**(3):686-702.
69. Timmermans MJ, Roelofs D, Marien J, van Straalen NM: **Revealing pancrustacean relationships: phylogenetic analysis of ribosomal protein genes places Collembola (springtails) in a monophyletic Hexapoda and reinforces the discrepancy between mitochondrial and nuclear DNA markers.** *BMC Evol Biol* 2008, **8**:83.
70. Kjer KM, Honeycutt RL: **Site specific rates of mitochondrial genomes and the phylogeny of eutheria.** *BMC Evol Biol* 2007, **7**:8.
71. Jones M, Gantenbein B, Fet V, Blaxter M: **The effect of model choice on phylogenetic inference using mitochondrial sequence data: lessons from the scorpions.** *Mol Phylogenet Evol* 2007, **43**(2):583-595.

## 7. Acknowledgements

I would like to thank all the people who supported me during my studies. They provided resources, ideas or knowledge. Without them this work would not have been possible.

First and foremost, I would like to thank my supervisor, PD Dr. Heike Hadrys, for giving me the opportunity to work on such an interesting project. I am thankful in many respects to her: the scientific and moral support, her advice and valuable suggestions and the opportunity and for giving me the opportunity to attend courses, meetings and field trips. I would also like to express my gratitude to Prof. Dr. Bernd Schierwater, who provided not only a wonderful working place and environment, but also gave me helpful suggestions in many respects.

I would like to thank Prof. Dr. Rob DeSalle, for his willingness to review this thesis and tolerating a long distance of travel.

Many colleagues who contributed to this work by providing samples, data or advice are listed in the acknowledgment section of each manuscript. Of these, I would like especially mention Karen Meusemann (Bonn) and Dr. Albert Melber (Hannover). I am thankful to my colleagues of the ITZ team in Hannover for their help, cordiality and friendship. Thanks to you all. I am especially grateful to Sandra Damm for many helpful discussions and all kinds of support. Thanks go also to Jutta Bunnenberg, Ann Kathrin Ketelsen and Angie Faust for their help in the laboratory, administration and linguistic help. I would also like to thank Björn Seegebarth and Marc Frederic Simon for their patience in helping me resolving every kind of computer problem.

I am also grateful for the financial support from the Boehringer Ingelheim Fonds for the Workshop on Molecular Evolution and Extended Sessions at the Marine Biological Laboratories (MBL), Woods Hole, USA, July/August 2008 and the financial support from the Gesellschaft der Freunde der Tierärztlichen Hochschule Hannover for the Annual Meeting of the Society for Molecular Biology and Evolution, Barcelona, Spain, 5.-8. June 2008.

Sven Sagasser I thank for his love and his boundless support. His motivation and continuous belief in me have made all this possible.

Without the continuous support of my parents, my brother and sister, I would have never come this far. Thank you for everything.

This work was financed by the special priority program “Deep Metazoan Phylogeny“ SPP 1174 of the German Science Foundation (DFG), grant no. HA 1947/5.

## 8. Curriculum Vitae

### Sabrina Simon

Stiftung Tierärztliche Hochschule Hannover  
ITZ, Ecology & Evolution  
Bünteweg 17d  
D-30559 Hannover, Germany  
+49-511-953-8402  
sabrina.simon@ecolevol.de

private:  
Podbielskistrasse 7  
D-30163 Hannover, Germany  
+49-174-3083711

---

#### PERSONAL DATA

Date of Birth: January, 6<sup>th</sup> 1981  
Place of Birth: Hannover, Germany  
Gender: Female  
Family Status: unmarried

#### EDUCATION

1991 – 2000      Graduation from Freihof-Gymnasium Göppingen with Allgemeine Hochschulreife

#### VOLUNTARY WORK

2000 – 2001      Volunteer work in the social sector, AWO Kreisverband Göppingen e.V.

#### UNIVERSITY EDUCATION

2001 – 2006      *Gottfried Wilhelm Leibniz University, Hannover, Germany*

- Undergraduate studies in Biological Science  
Major priority: ZOOLOGY, ECOLOGY, EVOLUTIONARY BIOLOGY

2006 – 2007      *Stiftung Tierärztliche Hochschule Hannover, Hannover, Germany*

- Diploma Thesis, Title: DEEP MOLECULAR PHYLOGENY OF THE PTERYGOTA: NEW MARKERS, NEW INSIGHTS, Advisor: Prof. Dr. rer. nat. Bernd Schierwater

02/2007 – present      *Stiftung Tierärztliche Hochschule Hannover, Hannover, Germany*

- Ph.D. in Biological Science, Title: DEEP MOLECULAR PHYLOGENY OF THE PTERYGOTA: Advisor: PD Dr. rer. nat. habil. Heike Hadrys



**SCIENTIFIC TRIPS**

- *Department of Molecular and Cellular Developmental Biology, Yale University, New Haven/CT, U.S.A., February-March 2007*  
Long-range PCR, comparative mitochondrial genomes analyses of basal pterygote species
- *Max Planck Institute for Molecular Genetics, Berlin, Germany, June 2007*  
Construction of cDNA-libraries for EST's of basal pterygotes
- *Crau d'Arles (Southern France), June 2008 & June 2009*  
Field work for sample collection

**ATTENDED WORKSHOPS**

- Course "Programming for Biology", October 12 - 27, 2009, Cold Spring Harbor Laboratory (CSH), Cold Spring Harbor, NY, USA
- Workshop on Molecular Evolution and Extended Topics Session, July 27 - August 15, 2008, Marine Biological Laboratories (MBL), Woods Hole, MA, USA

**TEACHING EXPERIENCE**

- Since 2006 Teaching Assistant: Intensive Course in Molecular Ecology and Evolution, ITZ, Stiftung Tierärztliche Hochschule Hannover
  - Laboratory methods and computational analyses of molecular evolution
- June 2008, June 2009: Teaching Assistant: Ecological Excursion (Crau&Camargue, Southern France), ITZ, Stiftung Tierärztliche Hochschule Hannover

**TRAVEL GRANT AWARDS**

- Travel Grant for the Annual Meeting of the Society for Molecular Biology and Evolution (SMBE), 5.-8. June 2008, Barcelona, Spain given by the Society of Friends of the Stiftung Tierärztliche Hochschule Hannover
- Travel Grant for the Workshop on Molecular Evolution and Extended Topics Session, July/August 2008, Marine Biological Laboratories (MBL), Woods Hole, MA, USA given by the Boehringer Ingelheim Fonds

**MEMBERSHIPS**

"Young Systematists" (workgroup of the European "Society of Biological Diversity")  
Society of Molecular Biology and Evolution

## 9. List of Publications

### Articles

- von Reumont BM, Meusemann K, Szucsich NU, Dell'Ampio E, Gowri-Shankar V, Bartel D, Simon S, Letsch HO, Stocsits RR, Luan Y, Wägele JW, Pass G, Hadrys H and B Misof (2009). Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships, *BMC Evolutionary Biology*, 9:119
- Simon S, Strauss S, von Haeseler A and H Hadrys (2009). A phylogenomic approach to resolve the basal pterygote divergence, *Molecular Biology and Evolution*, 26(12):2719-2730
- Simon S, Schierwater B and H Hadrys (2010). On the value of Elongation factor-1 $\alpha$  for reconstructing Pterygote insect phylogeny, *Molecular Phylogenetics and Evolution*, 54(2):651-656
- Meusemann K, von Reumont BM, Simon S, Roeding F, Kück P, Strauss S, Ebersberger I, Walz M, Pass G, Breuers S, Achter V, von Haeseler A, Burmester T, Hadrys H, Wägele JW and B Misof (2010). A phylogenomic approach to resolve the arthropod tree of life, *Molecular Biology and Evolution* (Epub 2010 June 9)
- Hadrys H, Simon S, Kaune B, Khadjeh S, Schmitt O, Schöner A and B Schierwater (2010). Isolation of Hox cluster genes from insects in development and evolution, *JEZ Part B: Molecular and Developmental Evolution*, in revision
- Simon S and H Hadrys (2010). Comprehensive analysis of nuclear rRNA genes to infer Insect phylogeny, prepared for submission to *BMC Evolutionary Biology*
- Simon S and H Hadrys (2010). The mitochondrial genome of two palaeopterous representatives: *Baetis* sp. (Ephemeroptera) and *Boyeria irene* (Odonata) – a mitogenomic approach to resolve the Palaeoptera problem, in preparation for submission to *BMC Genome*

### Presentations

- Simon S and H Hadrys (2009). A phylogenomic approach changes the traditional view on the origin of winged insects, 102. Annual meeting of the German Zoological Society, 25.-28. September 2009, Regensburg, Germany, Abstract Booklet, Page 194
- Simon S and H Hadrys (2009). The origin of winged insects: Identification of phylogenetic informative proteins using comparative phylogenomics; Celebrating Darwin: From The Origin of Species to Deep Metazoan Phylogeny, 4.-6. March 2009, Berlin, Germany, Abstract Booklet, Page 43

- Simon S and H Hadrys (2008). New markers and bioinformatics tools to reconstruct the phylogeny of winged insects (Pterygota). Annual Meeting of the Society for Molecular Biology and Evolution (SMBE), 5.-8. June 2008, Barcelona, Spain, Meeting Program, Page 27.
- Simon S, Angermann J and H Hadrys (2008). Deep Molecular Phylogeny of the Pterygota. Systematics 2008, 7.-11. April 2008, Göttingen, Germany, Abstract Booklet, Page 322.
- Simon S and H Hadrys (2008). Pterygota Phylogeny; Meeting of the German Science Foundation, priority program “Deep Metazoan Phylogeny”, Bonn, Germany, 14./15. January 2008.
- Simon S (2007). Evaluation and Exploration of new molecular markers systems for reconstructing pterygote phylogeny; workgroup meeting of evolutionary biology, University of Bielefeld, Bielefeld, Germany, 23. November 2007.
- Simon S and H Hadrys (2007). Evaluation of Elongation factor-1 $\alpha$  as molecular marker for pterygote systematics. 3rd Dresden Meeting on Insect Phylogeny, 21.-23. September 2007, Dresden, Germany. Abstract Booklet, Page 15-16.
- Simon S and H Hadrys (2007). Deep molecular phylogeny of Pterygota: Evaluation and Exploration of new molecular markers; 1<sup>st</sup> North German Evolution and Development Symposium, Hannover, Germany, 5. July 2007.
- Simon S, Sagasser S, Schierwater B and H Hadrys (2007). Genomic sequence of a hemocyanin in a terrestrial insect. 9. Annual meeting of the European “Society of Biological Diversity”, 20.-23. Februar 2007, Vienna, Austria. Abstract Booklet, Page 153.
- Simon S, Sagasser S, Schierwater B and H Hadrys (2006). Hemocyanin in Mantophasmatodea: Restricted homology among basal pterygote orders?, 99. Annual meeting of the of German Zoological Society, 16.-20. September 2006, Münster, Germany. Abstract Booklet, Page 150.

**Supplementary Table 1** – EF-1 $\alpha$  analyzed species of 20 pterygote insect orders with origin. Species ordered according to infraclasses. ID = Abbreviations of species names used in this study.

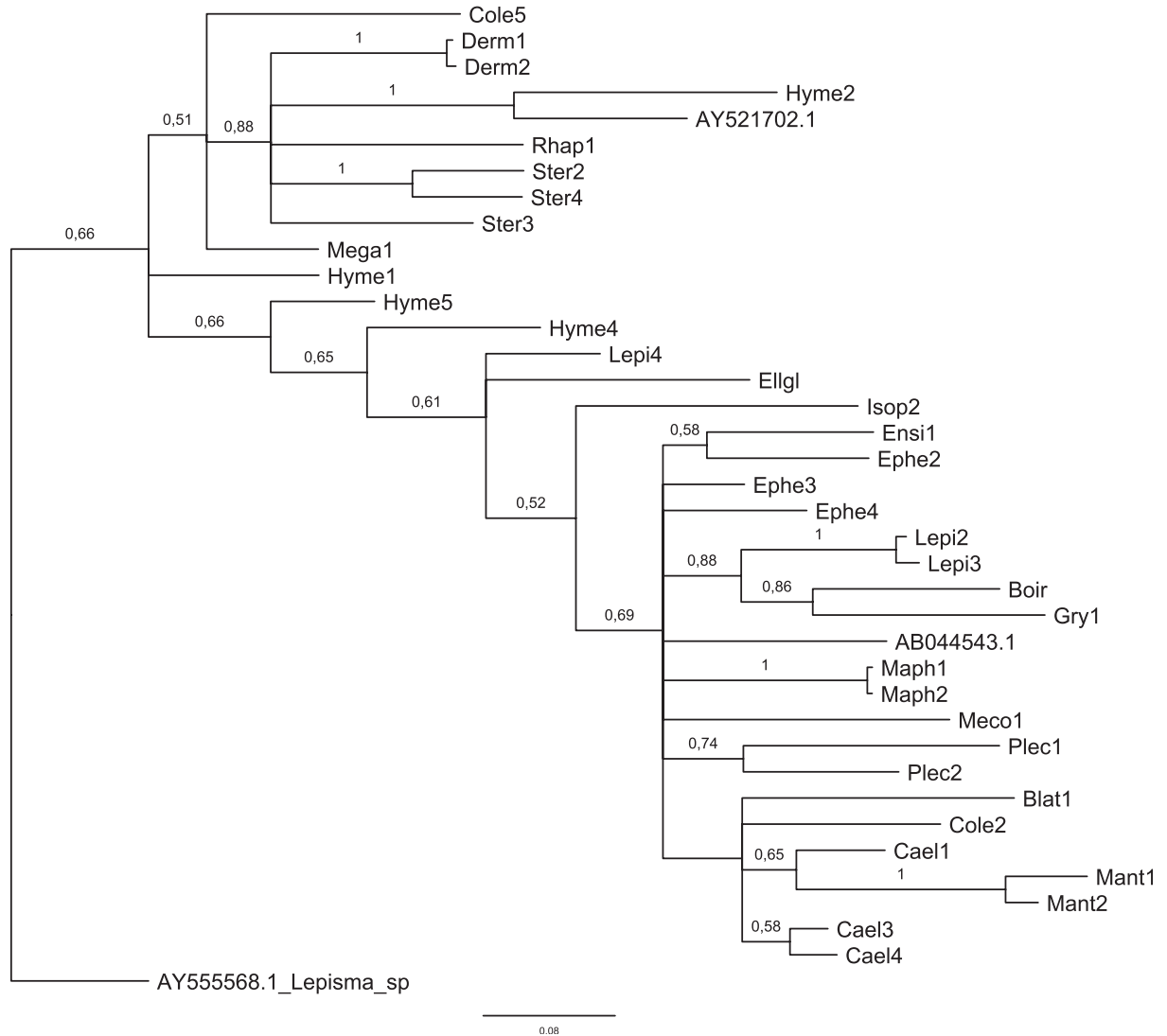
infraclass	order	species	origin	ID	accession nos.
Paleoptera	Ephemeroptera	<i>Ecdyonurus</i> sp.	Czechia	EPHE 2	EU414686
		<i>Baetis</i> sp.	Germany	EPHE 3	EU414687
		<i>Siphonura aestivalis</i> (EATON)	Czechia	EPHE 4	EU414688
	Odonata	<i>Elatoneura glauca</i>	Namibia	EII gl	EU414705
		<i>Boyeria irene</i>	France	Boir	EU414704
Polyneoptera	Plecoptera	<i>Isoperla</i> sp.	Czechia	PLEC 1	EU414708
		<i>Nemoura (cinerea)</i> (RETZ.)	Czechia	PLEC 2	EU414709
	Mantophasmatodea	<i>Mantophasma zephyra</i>	South Africa	MAPH 1	EU414700
		<i>Tyrannophasma gladiator</i>	South Africa	MAPH 2	EU414701
	Grylloblattodea	<i>Galloisiana yuasai</i> ASAHINA	Japan	GRY 1	EU414714
	Dermaptera	<i>Forficula auricularia</i> L.	Germany	DERM 1	EU414681
		<i>Apterygida media</i> (HAGENB.)	Germany	DERM 2	EU414682
	Mantodea	<i>Rhombodera</i> sp.	Germany (breeding)	MANT 1	EU414698
		<i>Hierodula</i>	Germany (breeding)	MANT 2	EU414699
	Blattodea	<i>Blaberus fusca</i> BRUNN.	South Africa (breeding)	BLAT 1	EU414675
	Isoptera	<i>Mastotermes darwiniensis</i>	Australia (breeding)	ISOP 2	EU414693
	Ensifera	<i>Achaeta domesticus</i> (L.)	Germany	ENSI 1	EU414685
	Caelifera	<i>Anacridium aegypticum</i> (L.)	Tunisia	CAEL 1	EU414676
		<i>Acrida turrita</i> (L.)	Tunisia	CAEL 3	EU414677
		<i>Sphingonotus</i> sp.	Tunisia	CAEL 4	EU414678
Paraneoptera	Sternorrhyncha	<i>Aphis</i> sp.	Germany	STER 2	EU414710
		<i>Lachnus</i> sp.	Germany	STER 4	EU414712
		<i>Pulvinaria regalis</i> CANARD	Germany	STER 3	EU414711
Holometabola	Coleoptera	<i>Dermestes maculatus</i> DE GEER	Germany	COLE 2	EU414679
		<i>Pimelia</i> sp.	Tunisia	COLE 5	EU414680
	Rhaphidioptera	<i>Phaestigma</i> sp.	Germany	RHAP 1	EU414713
	Megaloptera	<i>Sialis lutaris</i> (L.)	Germany	MEGA 1	EU414703
	Planipennia	gen. sp.	Germany	PLAN 3	EU414706
		<i>Euroleon nostras</i> (FOURCR.)	France	PLAN 4	EU414707
	Hymenoptera	<i>Nomada</i> sp.	Germany	HYME 1	EU414689
		<i>Bombus terrestris</i> (L.)	Germany	HYME 2	EU414690
		<i>Scolia</i> sp.	Tunisia	HYME 3	EU414691
		gen. sp.	Germany	HYME 5	EU414692
	Lepidoptera	<i>Anthocharis cardamines</i> L.	Germany	LEPI 1	EU414694
		<i>Pieris napi</i> L.	Germany	LEPI 2	EU414695
		<i>Pieris rapae</i> L.	Germany	LEPI 3	EU414696
		<i>Eurhypara hortulana</i> (L.)	Germany	LEPI 4	EU414697
	Mecoptera	<i>Boreus hyemalis</i> (L.)	Germany	MECO 1	EU414702
	Diptera	<i>Bombylius major</i> L.	Germany	DIPT 2	EU414683
		<i>Ctenophora</i> sp.	Germany	DIPT 3	EU414684

**Supplementary Table 2** – Histone H3 analyzed species of 19 pterygote insect orders with origin. Species ordered according to infraclasses. ID = Abbreviations of species names used in this study.

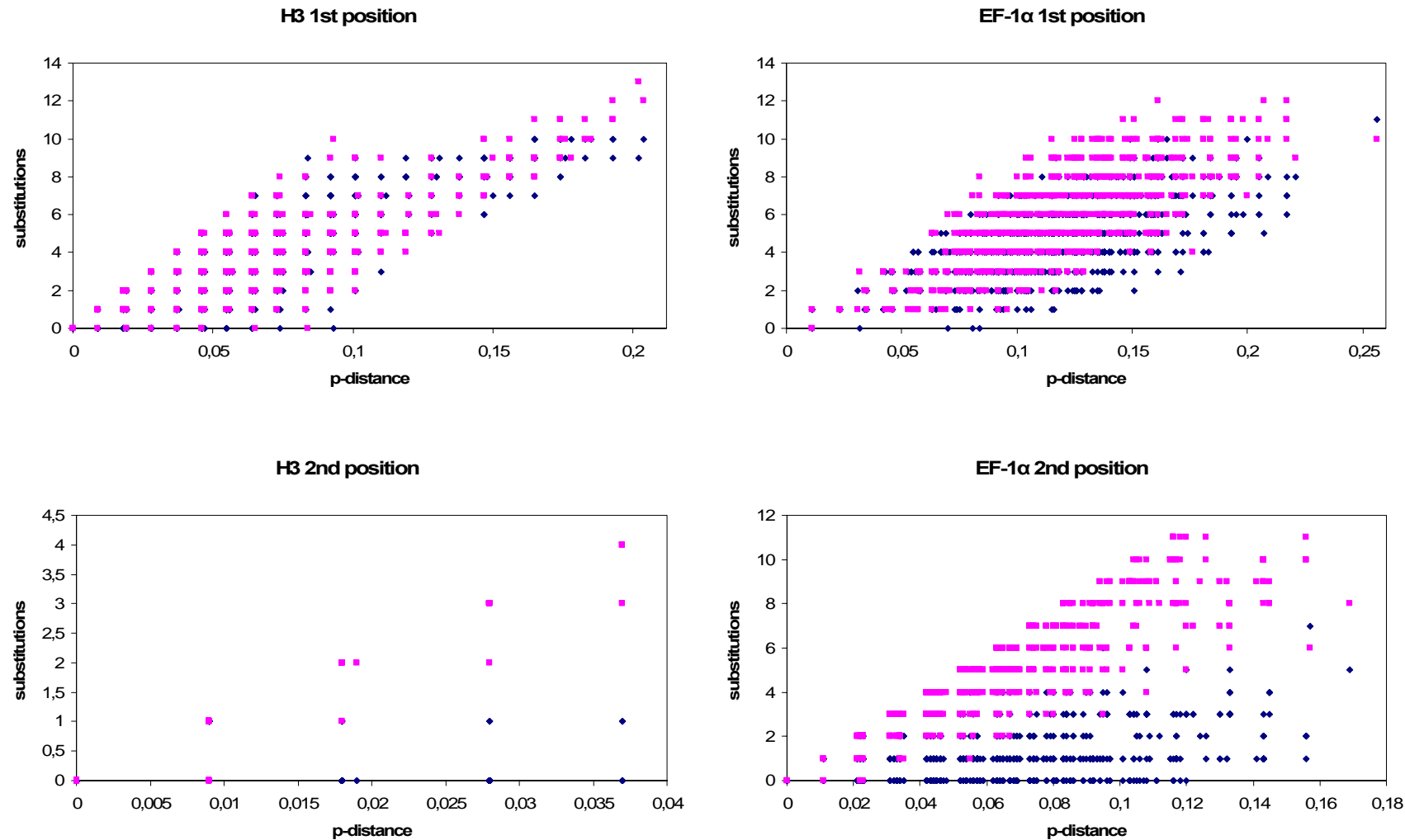
infraclass	order	species	origin	ID	accession nos.
Palaeoptera	Ephemeroptera	<i>Ecdyonurus</i> sp.	Czechia	EPHE 2	GU066908
		<i>Baetis</i> sp.	Germany	EPHE 3	GU066909
		<i>Siphonurus aestivalis</i> (EATON)	Czechia	EPHE 4	GU066910
	Odonata	<i>Elatoneura glauca</i>	Namibia	EII gl	GU066927
		<i>Boyeria irene</i>	France	Boir	GU066926
Polyneoptera	Plecoptera	<i>Isoperla</i> sp.	Czechia	PLEC 1	GU066928
		<i>Nemoura (cinerea)</i> (RETZ.)	Czechia	PLEC 2	GU066929
	Mantophasmatodea	<i>Mantophasma zephyra</i>	South Africa	MAPH 1	GU066922
		<i>Tyrannophasma gladiator</i>	South Africa	MAPH 2	GU066923
	Grylloblattodea	<i>Galloisiana yuasai</i> ASAHINA	Japan	GRY 1	GU066933
	Dermaptera	<i>Forficula auricularia</i> L.	Germany	DERM 1	GU066905
		<i>Apterygida media</i> (HAGENB.)	Germany	DERM 2	GU066906
	Mantodea	<i>Rhombodera</i> sp.	Germany (breeding)	MANT 1	GU066920
		<i>Hierodula</i>	Germany (breeding)	MANT 2	GU066921
	Blattodea	<i>Blaberus fusca</i> BRUNN.	South Africa (breeding)	BLAT 1	GU066899
	Isoptera	<i>Mastotermes darwiniensis</i>	Australia (breeding)	ISOP 2	GU066915
	Ensifera	<i>Achaeta domesticus</i> (L.)	Germany	ENSI 1	GU066907
	Caelifera	<i>Anacridium aegypticum</i> (L.)	Tunisia	CAEL 1	GU066900
		<i>Acrida turrita</i> (L.)	Tunisia	CAEL 3	GU066901
		<i>Sphingonotus</i> sp.	Tunisia	CAEL 4	GU066902
Paraneoptera	Sternorrhyncha	<i>Aphididae</i> sp.	Germany	STER 2	GU066930
		<i>Lachnus</i> sp.	Germany	STER 4	GU066932
		<i>Pulvinaria regalis</i> CANARD	Germany	STER 3	GU066931
Holometabola	Coleoptera	<i>Dermestes maculatus</i> DE GEER	Germany	COLE 2	GU066903
		<i>Pimelia</i> sp.	Tunisia	COLE 5	GU066904
	Rhaphidioptera	<i>Phaestigma</i> sp.	Germany	RHAP 1	GU066917
	Megaloptera	<i>Sialis lutaris</i> (L.)	Germany	MEGA 1	GU066925
	Hymenoptera	<i>Nomada</i> sp.	Germany	HYME 1	GU066911
		<i>Bombus terrestris</i> (L.)	Germany	HYME 2	GU066912
		<i>Tenthredinidae</i> gen sp.	Tunisia	HYME 4	GU066913
		<i>Cynipidae</i> gen. sp.	Germany	HYME 5	GU066914
		<i>Pieris napi</i> L.	Germany	LEPI 2	GU066916
		<i>Pieris rapae</i> L.	Germany	LEPI 3	GU066918
		<i>Eurrhynx hortulata</i> (L.)	Germany	LEPI 4	GU066919
	Mecoptera	<i>Boreus hyemalis</i> (L.)	Germany	MECO 1	GU066924
	Diptera	<i>Drosophila lutescens</i>			AB044543
		<i>Dolichopeza subalbipes</i>			AY521702

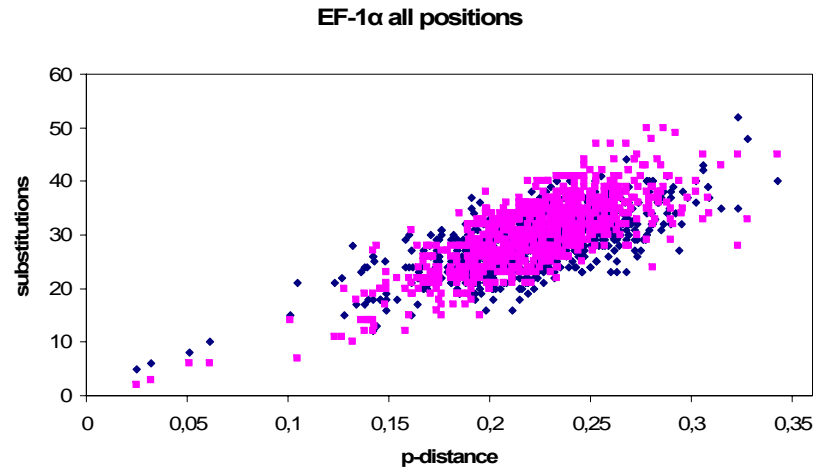
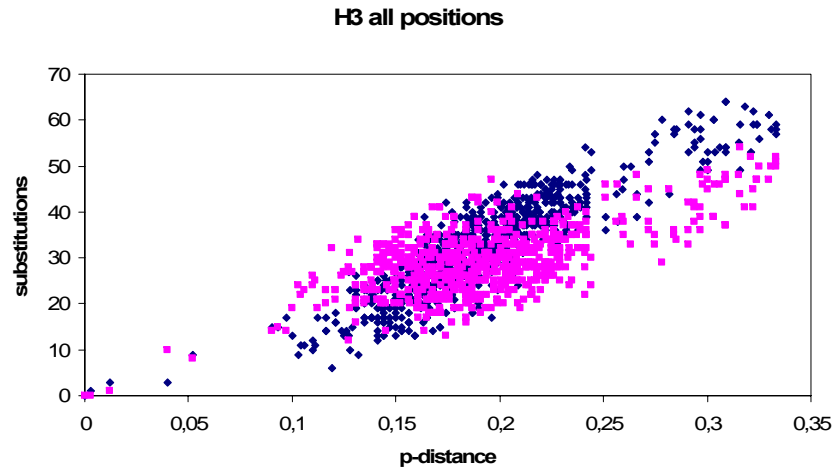
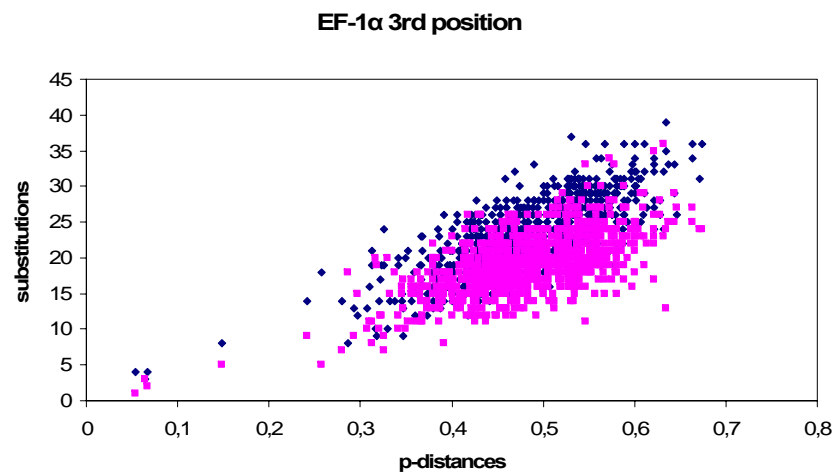
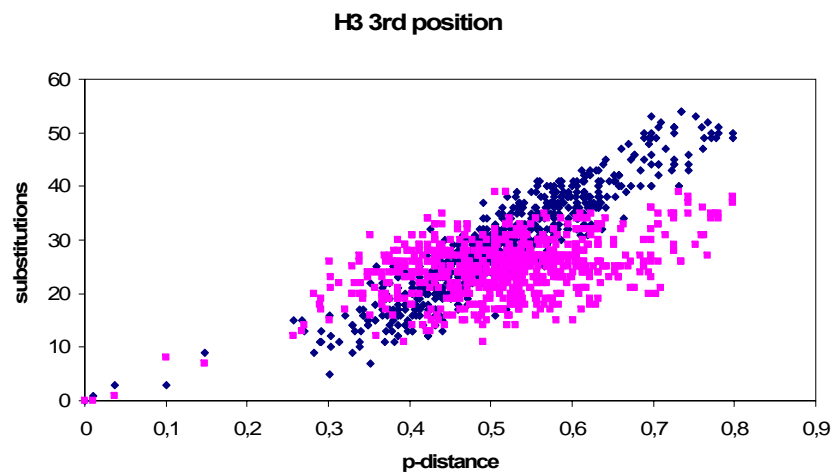
Note: The only differences between EF-1 $\alpha$  and Histone H3 analyzed species are: Hyme3 (EF-1 $\alpha$ ) / Hyme4 (Histone H3); Plan3 and Plan4 missing for Histone H3; Dipteran species vary between EF-1 $\alpha$  and Histone H3.

**Supplementary Fig. 1** – Phylogeny of Pterygota based on phylogenetic analyses of Histone 3 nucleotide sequence. Majority rule consensus of topologies generated via MrBayes with 3,000,000 generations (first 7500 trees were discarded as “burn-in”) using the GTR+SSR model. Numbers at nodes represent percentage of group inclusion among all topologies generated with MrBayes using the GTR+SSR model. The sequence of the apterygote *Lepisma* sp. (Thysanura: GenBank accession no. **AY555568**) was used as the outgroup.



**Supplementary Fig. 2** – Substitution pattern of transitions (blue) and transversions (pink) for Histone 3 and Elongation factor-1 $\alpha$  for all data and each codon position. The number of transitions and transversions is plotted against total sequence divergence (uncorrected  $p$  distance) for all pairwise comparison of taxa.



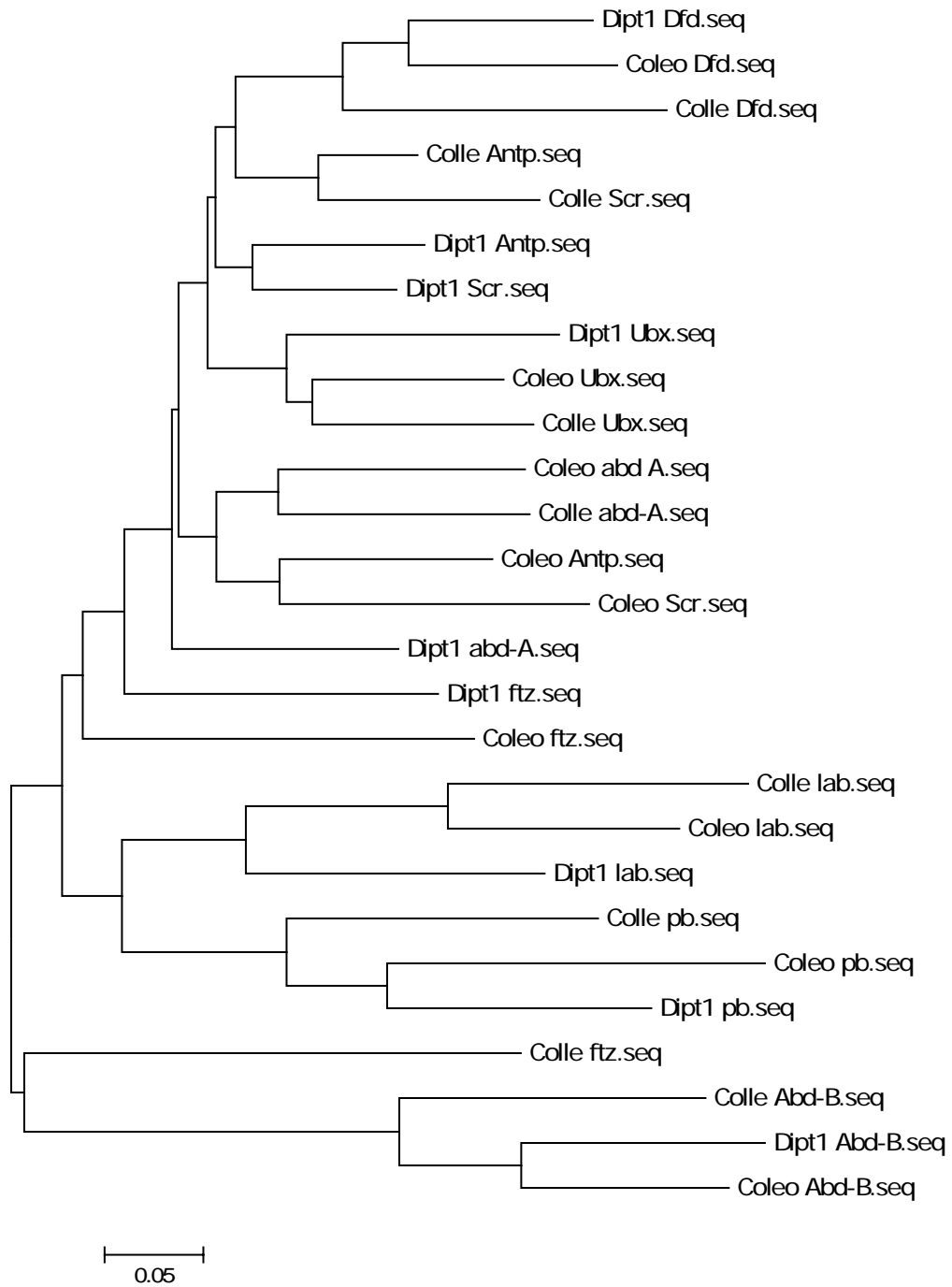




**Table S1:** Taxa list and GenBank accession numbers.

<b>species/gene</b>	<b>lab</b>	<b>pb</b>	<b>Dfd</b>	<b>Scr</b>	<b>ftz</b>	<b>Antp</b>	<b>Ubx</b>	<b>abd-A</b>	<b>Abd-B</b>
<i>Artemia franciscana</i>		AF363018	X70078	X70080		AF435786	X70081	X70076	
<i>Thermobia domestica</i>	AF104008	AF104009	AF104005	AF104010	AY456923	AF104003		AF104001	AF104002
<i>Schistocerca gregaria</i>	AF363015		AF363016	X73981		U32943	AF363017	X54674	X69161
<i>Gryllus bimaculatus</i>				AB194276		AB194275	AB194278		
<i>Apis mellifera</i>	XM_001120278	XM_394125	XM_001120045	XM_394121		NM_001011571	XM_623986	XM_394120	XM_394119
<i>Nasonia vitripennis</i>	XM_001603789	XM_001603738		AY684807	XM_001603620	XM_001602672	XM_001603571		XM_001603544
<i>Bombyx mori</i>			D83534	D83533		D16684	X62618	NM_001114159	X62619
<i>Manduca sexta</i>						U63301	U63300	S77989	
<i>Tricolium castaneum</i>	AF231104	AF187068	U81039	AF227628	U14732	AF228509	AF146649	AF017415	AF227923
<i>Anopheles gambiae</i>	AF269153	AF269154	AF269155	AF080564		AF080565	AF080563	AF080566	DQ383819
<i>Aedes aegypti</i>	XM_001650461		XM_001660448			XM_001660446		X67132	
<i>Drosophila melanogaster</i>	NM_057265	X63728	X05136	X05228	X00854	M20705	X76210	X54453	X16134

**Figure S1:** Neighbor-Joining tree of nine complete homeobox sequences from *Folsomia candida* (Colle), *Drosophila melanogaster* (Dipt) and *Tribolium castaneum* (Coleo)



## Additional file 1: Taxa list.

Order	Taxon	Accession number 28S rRNA	length 28S rRNA (bp)	Accession number 18S rRNA	length 18S rRNA (bp)
Arachnida	<i>Amblyomma americanum</i>	AF291874	4005	AF291874	1815
	<i>Dermacentor</i> sp. *	AY859582	3920	L76340	1784
	<i>Chaloeiridius</i> cf. <i>termitophilus</i>	AY859558	3394	AY859559	1773
	<i>Pandinus imperator</i>	AY210830	3777	AY210831	1762
	<i>Siro rubens</i>	AY859602	3762	U36998	1809
	<i>Eremobates</i> sp.	AY859572	3833	AY859573	1767
	<i>Aphonopelma hentzi</i> *	AY210803	3819	DQ639776	1750
	<i>Misumenops asperatus</i>	AY210461	3467	AY210445	1786
	<i>Mastigoproctus giganteus</i>	AY859587	3796	AF005446	1790
	<i>Paraphrynos</i> sp.	AY859594	3785	AF005445	1777
	<i>Limulus polyphemus</i>	AF212167	3772	L81949	1807
	<i>Callipallene</i> sp.	AY210807	3900	AF005439	1817
	<i>Colossendeis</i> sp.	EU420133 ** (v. Reumont)	3864	EU420135 ** (v. Reumont)	1798
	<i>Anoplodactylus portus</i>	AY859550	3893	AY859551	1809
	<i>Nymphon stroemii</i>	EU420134 ** (v. Reumont)	3818	EU420136 ** (v. Reumont)	1825
	<i>Artemia</i> sp. *	AY210805	3628	AJ238061	1809
Anostraca	<i>Triops cancriformis</i>	EU370435 ** (v. Reumont)	3420	EU370422 ** (v. Reumont)	1784
	<i>Triops longicaudatus</i>	AY157606	3458	AF144219	1809
Notostraca	<i>Daphnia</i> cf. <i>magna</i>	EU370436 ** (v. Reumont)	3823	EU370423 ** (v. Reumont)	2291
	<i>Bosmina</i> sp. *	EU370437 ** (v. Reumont)	3332	Z22731	1875
Diplostraca	<i>Eulirnadia texana</i>	AY859574	3665	AF144211	1813
	<i>Heterocypris incongruens</i>	EU370438 ** (v. Reumont)	3279	EU370424 ** (v. Reumont)	1786
Ostracoda	<i>Pontocypris mytiloides</i>	EU370439 ** (v. Reumont)	3672	EU370425 ** (v. Reumont)	1897
	<i>Semibalanus balanoides</i>	EU370440 ** (v. Reumont)	3274	EU370426 ** (v. Reumont)	1847
Cirripedia	<i>Megabalanus californicus</i>	AY859588	3720	AY520632	1812
	<i>Pollicipes pollicipes</i>	EU370441 ** (v. Reumont)	3549	EU370427 ** (v. Reumont)	1852
Branchiura	<i>Argulus foliaceus</i>	EU370442 ** (v. Reumont)	3512	EU370428 ** (v. Reumont)	1851
	<i>Derocheilicaris typicus</i>	EU370443 ** (v. Reumont)	3663	EU370429 ** (v. Reumont)	2171
Mystacocarida	<i>Cyclopidae</i> sp. *	AY210813	3536	AJ746334	1808
	<i>Chondracanthus lophii</i>	DQ180341	3465	L34046	1810
Copepoda	<i>Tigriopus</i> cf. <i>fulvus</i>	EU370444 ** (v. Reumont)	3532	EU370430 ** (v. Reumont)	1792
	<i>Canuella perplexa</i>	EU370445 ** (v. Reumont)	3462	EU370432 ** (v. Reumont)	1573
Remipedia	<i>Lepeophtheirus salmonis</i>	DQ180342	3692	AF208263	1799
	<i>Speleonectes tulumensis</i>	EU370446 ** (v. Reumont)	3797	EU370431 ** (v. Reumont) / L81936	1302 / 1965
Cephalocarida	<i>Hutchinsoniella macracantha</i>	EF189645	2480	L81935	2018
	<i>Nebalia</i> sp.	EU370447 ** (v. Reumont)	3519	EU370433 ** (v. Reumont)	1789
Leptostraca	<i>Anaspides tasmaniae</i>	AY859549	3997	L81948	1827
	<i>Heteromysis</i> sp.	AY859578	3400	AY743946	1724
Anaspidacea	<i>Homarus americanus</i>	AY859581	4351	AY743945	1758
	<i>Penaeus vannamei</i> *	AF124597	5820	DQ079766	1781
Decapoda	<i>Squilla empusa</i>	AY210842	3913	L81946	1817
	<i>Raillietiella</i> sp. *	EU370448 ** (v. Reumont) / AY744894	1286 / 1983	EU370434 ** (v. Reumont)	1814
Pentastomida	<i>Craterostigma tasmanianus</i>	EU376009 ** (Bartel)	4024	EU368617 ** (Meusemann)	1786
	<i>Otostigma politus</i>	DQ666180	4170	DQ666177	1868
Chilopoda	<i>Scolopendra mutilans</i>	DQ666181	4174	DQ666178	1848
	<i>Scutigera coleoptrata</i>	AY859601	4024	AF000772	1865
Diplopoda	<i>Lithobius forficatus</i>	EF199984	3913	EU368618 ** (Meusemann)	1752
	<i>Polyxenus lagurus</i>	EU376011 ** (Bartel)	3967	EU368619 ** (Meusemann)	1733
Protura	<i>Monographtus</i> sp.	EF192437 ** (Bartel / Luan)	3866	AY596371	1744
	<i>Paradoxosomatidae</i> sp.	DQ666182	4288	DQ666179	1797
Pauropoda	<i>Polydesmus complanatus</i>	EU376010 ** (Bartel)	4271	EU368620 ** (Meusemann)	1689
	<i>Cherokia georgiana</i>	AY859562	4225	AY859563	1781
Symphyla	<i>Orthoporus</i> sp.	AY210828	4124	AY210829	1791
	<i>Cylindroiulus caeruleocinctus</i>	EF199985	4084	EU368621 ** (Meusemann)	1753
Symphylla	<i>Allopaupropus</i> sp.	DQ666185	4406	DQ399857	2227
	<i>Paupropodidae</i> sp.	EU376012 ** (Bartel)	4238	EU368622 ** (Meusemann)	2250
Protura	<i>Scutigera</i> sp.	DQ666184	4471	DQ399856	1902
	<i>Hansenella</i> sp.	AY210821-22	4539	AY210823	1925
Diplura	<i>Symphylla</i> sp.	DQ666183	4558	DQ399855	2057
	<i>Acerentomon franzi</i>	EF199976	4099	EU368597 ** (Meusemann)	1790
Collembola	<i>Baculentulus densus</i> *	EU376049	4100	AY037169	1984
	<i>Eosentomon</i> sp.	EU376047 ** (Dell'Ampio)	3654	EU368598 ** (Meusemann)	1860
Diplura	<i>Eosentomon sakura</i>	EF192434 ** (Dell'Ampio / Luan)	3789	AY596355	1948
	<i>Stenotomon erythranum</i>	EF192442 ** (Dell'Ampio / Luan)	4043	AY596358	1934
Diplura	<i>Campodeidae</i> sp.	AY859560	3718	AY859561	1866
	<i>Campodea augens</i>	EF199977	4010	EU368599 ** (Meusemann)	1788
Collembola	<i>Lepidocampa weberi</i>	EU376050	4061	AY037167	1878
	<i>Cataglyphis aquilonaris</i>	EF199978	5016	EU368600 ** (Meusemann)	2154
Collembola	<i>Parajapyx emeryanus</i>	EF192440 ** (Dell'Ampio / Luan)	4143	AY037168	2120
	<i>Ocostigma sinensis</i>	EF192439 ** (Dell'Ampio / Luan)	4001	AY145134	2138
Collembola	<i>Tetradontophora bielensis</i>	EU376051	3868	AY555519	1760
	<i>Gomphiocephalus hodgsoni</i>	EF199969	3893	EU368601 ** (Meusemann)	1746
Collembola	<i>Triacanthella</i> sp.	AY859609	3823	AY859610	1758
	<i>Bilobella aurantiaca</i>	AJ251729	3934	EU368602 ** (Meusemann)	1759
Collembola	<i>Anurida maritima</i>	AJ251738	3965	EU368603 ** (Meusemann)	1680
	<i>Podura aquatica</i>	EF199970	3899	EU368604 ** (Meusemann)	1696
Collembola	<i>Cryptopygus antarcticus</i>	EF199971	3862	EU368605 ** (Meusemann)	1724
	<i>Isotoma viridis</i>	EU376052	3866	AY596361	1748
Collembola	<i>Orchesella villosa</i>	EF199972	3867	EU368606 ** (Meusemann)	1739
	<i>Pogonognathellus flavescens</i>	EU376053	3874	EU368607 ** (Meusemann)	1688
Collembola	<i>Megalothorax minimus</i>	EF199975	3868	EU368608 ** (Meusemann)	1703
	<i>Sminthurus viridis</i>	EF199973	3912	EU368609 ** (Meusemann)	1695
Collembola	<i>Allacma fusca</i>	EU376054	3877	EU368610 ** (Meusemann)	1759
	<i>Dicyrtomina saundersi</i>	EF199974	3871	EU368611 ** (Meusemann)	1739
Collembola	<i>Machilis hrabei</i>	EF199981	3750	EU368612 ** (Meusemann)	1703
	<i>Lepismachilis y-signata</i>	EF199980	3826	EU368613 ** (Meusemann)	1679
Collembola	<i>Pedetontus okajimae</i>	EU376055	3800	EU368614 ** (Meusemann)	1742
	<i>Lepisma saccharina</i>	EU376048 ** (Dell'Ampio)	3506	EU368615 ** (Meusemann)	1703
Collembola	<i>Ctenolepisma longicaudata</i>	AY210810	3907	EU368616 ** (Meusemann)	1744
	<i>Brachytroch pratense</i>	EU424323 ** (Letsch)	3738	AF461232	1737
Collembola	<i>Aeshna juncea</i>	EU424324 ** (Letsch)	3736	AF461231	1767
	<i>Oxygastra curtisi</i>	EU424325 ** (Letsch)	3736	DQ008194	1787
Collembola	<i>Cordulia aenea</i>	EU424326 ** (Letsch)	3795	AF461236	1768
	<i>Somatochlora flavomaculata</i>	EU424327 ** (Letsch)	3795	AF461242	1757
Collembola	<i>Epiphlebia superstes</i>	EU424328 ** (Letsch)	3736	AF461247	1835
	<i>Progomphus obscurus</i>	EU424329 ** (Letsch)	3756	AY749909	1843
Collembola	<i>Sympetrum danae</i>	EU424330 ** (Letsch)	3756	AF461243	1754
	<i>Leucorhinia</i> sp.	AY859583	4114	AY859584	1815
Collembola	<i>Lestes viridis</i>	EU424331 ** (Letsch)	3747	AJ421949	1867

### Appendix 3: Major arthropod relationships inferred from rRNA genes

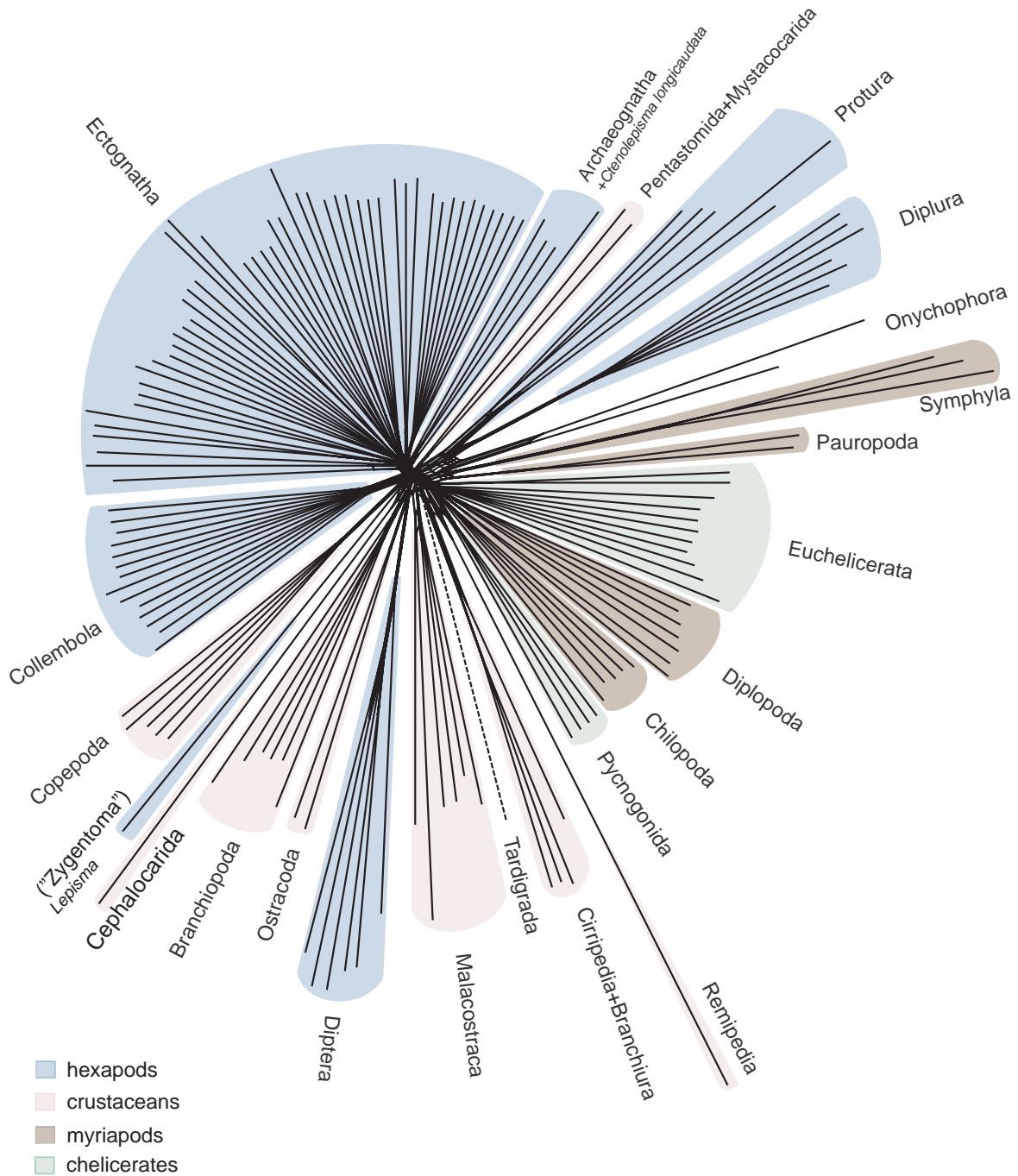
Order	Taxon	Accession number 28S rRNA	length 28S rRNA (bp)	Accession number 18S rRNA	length 18S rRNA (bp)
Ephemeroptera	<i>Callibaetis ferrugineus</i>	AY859557	3887	AF370791	1812
	<i>Epeorus sylvicola</i> *	EU414715 ** (Simon)	3680	AY749837	1808
Phasmatodea	<i>Siphonura aestivalis</i> *	EU414716 ** (Simon)	4151	DQ008181	1784
	<i>Carausius morosus</i>	EU426878 ** (Simon)	3737	X89488	1899
Mantophasmatodea	<i>Bacillus rossius</i>	EU426879 ** (Simon)	3889	AY121180	1891
	<i>Mantophasma zephyra</i> *	EU414719 ** (Simon)	3383	DQ874153	2018
Mantodea	<i>Tyrannophasma gladiator</i>	EU426875 ** (Simon)	3878	AY521863	2074
	<i>Mantis religiosa</i>	AY859585	3990	AY491153	1734
Blattaria	<i>Hierodula membranacea</i> *	EU414720 ** (Simon)	3603	AY491194	1734
	<i>Gromphadorhina laevigata</i>	AY210819	4015	AY210820	1877
Isoptera	<i>Ectobius lapponicus</i>	EU426877 ** (Simon)	4006	DQ874125	1808
	<i>Blattella germanica</i>	AF005243	3931	AF005243	1964
Dermaptera	<i>Zootermopsis angusticollis</i>	AY859614	4183	AY859615	1873
Plecoptera	<i>Forficula auricularia</i>	EU426876 ** (Simon)	4016	Z97594	1873
Hemiptera	<i>Isoperla</i> sp. *	EU414717 ** (Simon)	4299	AF461256	2054
	<i>Nemoura flexuosa</i> *	EU414718 ** (Simon)	3256	AF461257	1763
	<i>Pyrrhocoris apterus</i> *	EU414725 ** (Simon)	3389	AY627318	1829
	<i>Rhaphigaster nebulosa</i>	EU426880 ** (Simon)	3983	X89495	1924
	<i>Harpocera thoracica</i> *	EU414726 ** (Simon)	3405	AY252388	1895
	<i>Cercopis vulnerata</i> *	EU414724 ** (Simon)	3615	AY744798	1856
Orthoptera	<i>Clastoptera obtusa</i>	AF304569	3201	AY744784	1859
	<i>Pectinariophyes reticulata</i>	AF304570	3259	AY744778	1848
	<i>Gomphocerinae</i> sp.	AY859546	4187	AY859547	1864
	<i>Anacridium aegyptium</i> *	EU414723 ** (Simon)	3819	AY379759	1833
	<i>Acheta domesticus</i>	AY859544	4092	X95741	1802
	<i>Leptophyes punctatissima</i> *	EU414721 ** (Simon)	3918	AY521867	1897
Hymenoptera	<i>Pholidoptera griseoptera</i> *	EU414722 ** (Simon)	3950	Z97587	1884
	<i>Myrmecia croslandi</i>	AB052895	3460	AB121786	1766
	<i>Vespula pensylvanica</i>	AY859612	3912	AY859613	1871
	<i>Nomada</i> sp. *	EU414727 ** (Simon)	3386	AY703484	1854
Coleoptera	<i>Scolia</i> sp. *	EU414728 ** (Simon)	3405	EF012932	1851
	<i>Tenthredinidae</i> sp. *	EU414729 ** (Simon)	3472	AF423781	1836
Siphonaptera	<i>Tenebrio</i> sp. *	AY210843	4459	X07801	2083
	<i>Silpha obscura</i>	EU426881 ** (Simon)	2783	AJ810737	1930
Mecoptera	<i>Ctenocephalides felis</i> *	EU414732 ** (Simon)	3333	AF423914	1878
	<i>Meropie tuber</i>	DQ202351	3736	AF286287	1886
	<i>Boreus hyemalis</i>	EU426882 ** (Simon)	3534	AF423882	1881
Lepidoptera	<i>Pieris napi</i> *	EU414731 ** (Simon)	3743	AF423785	1856
Trichoptera	<i>Oxyethira rossi</i> *	DQ202352	3869	AF423801	1848
Diptera	<i>Trienodes</i> sp. *	EU414730 ** (Simon)	3095	AF286300	1897
	<i>Acricotopus lucens</i>	AJ586562	3910	AJ586561	1939
	<i>Chironomus tentans</i>	X99212	3973	X99212	1528
	<i>Anopheles albimanus</i>	L78065	4022	L78065	1977
	<i>Aedes albopictus</i>	L22060	4102	X57172	1950
	<i>Drosophila melanogaster</i>	M21017	3900	M21017	1995
Onychophora	<i>Simulium sanctipauli</i>	AF403805	3733	AF403800	1912
	<i>Peripatus</i> sp.	AY210836	3297	AY210837	2476
Tardigrada	<i>Peripatoides novaezealandiae</i>	AF342793	4570	AF342794	2064
	<i>Milnesium</i> sp. *	AY210826	3579	U49909	1844

\* indicates concatenated 18S and 28S rRNA sequences from different species. For combinations of genes to construct concatenated sequences of chimeran taxa, see Additional file 5.

\*\* contributed sequences in the present study (author of sequences).

**Additional file 2**

LogDet corrected network of concatenated 18S and 28S rRNA alignment. LogDet corrected network plus invariant site models (30.79% invariant sites) using SplitsTree4 based on the concatenated 18S and 28S rRNA alignment after exclusion of randomly similar sections evaluated with ALISCORE.



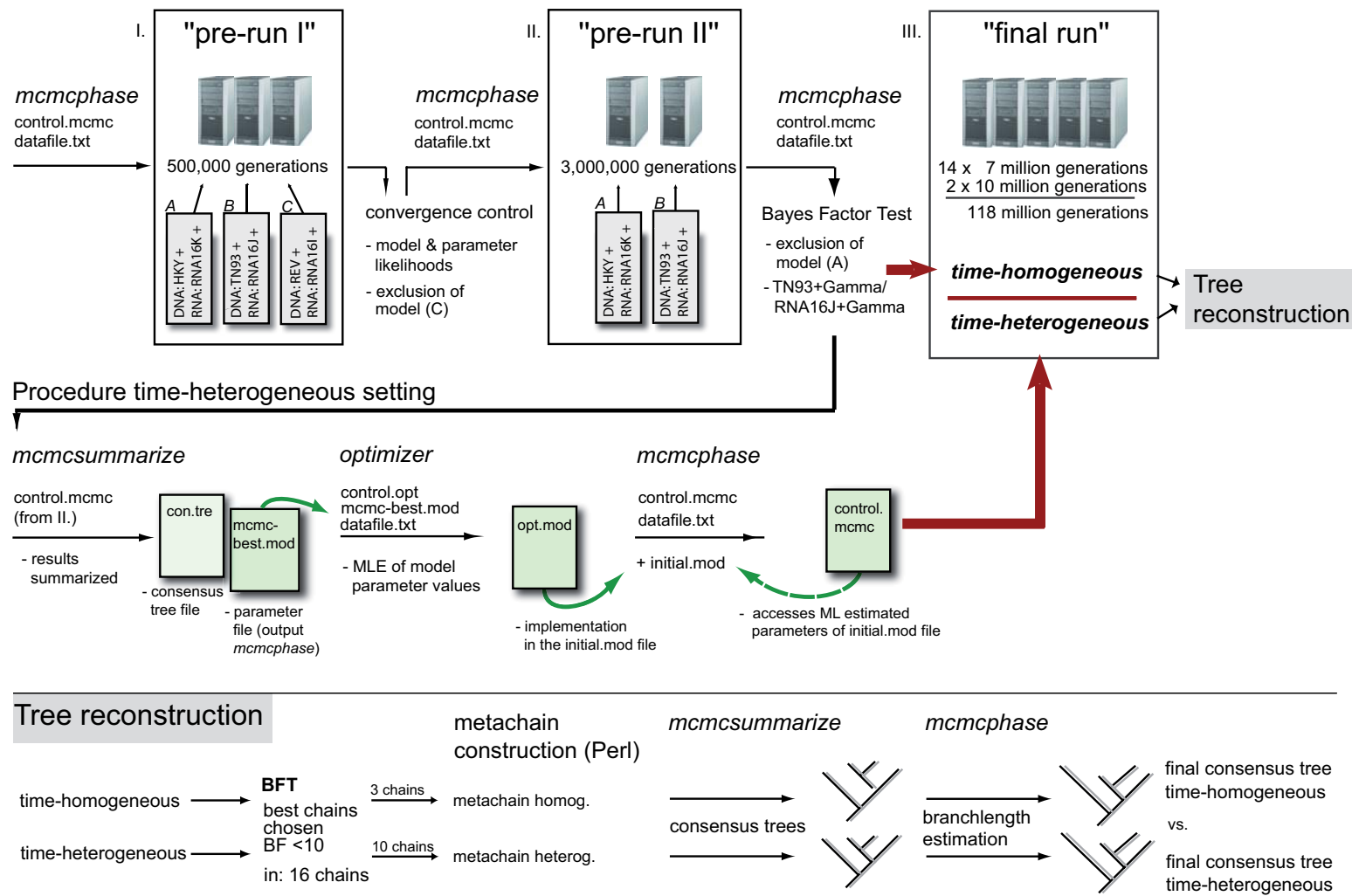
## Additional file 3

Bayesian support values for selected clades.

Clade	pP: time-heterogeneous	pP: time-homogeneous
Symphyla	1.0	1.0
(Paupoda, Onychophora)	0.97	1.0
Paupoda	1.0	1.0
Onychophora	1.0	1.0
Chelicerata	0.91	1.0
Pycnogonida	1.0	1.0
Euchelicerata (without Pycnogonida)	0.89	1.0
Myriapoda partim (excl. Symphyla & Paupoda): (Diplopoda, Chilopoda)	0.97	0.98
Diplopoda	0.99	1.0
Chilopoda	1.0	1.0
Myriochelata partim: ((Diplopoda, Chilopoda)(Euchelicerata, Pycnogonida))	0.97	1.0
(Myriochelata partim, Pancrustacea)	0.95	0.98
Pancrustacea	1.0	1.0
(( <i>Derocheilocaris</i> , Ostracoda)( <i>Speleonectes</i> ( <i>Argulus</i> , Cirripedia)))	0.33	-
( <i>Derocheilocaris</i> , Ostracoda)	0.62	-
(( <i>Derocheilocaris</i> , <i>Raillietiella</i> )( <i>Speleonectes</i> ( <i>Argulus</i> , Cirripedia)))	-	0.59
( <i>Derocheilocaris</i> , <i>Raillietiella</i> )	-	0.75
( <i>Speleonectes</i> ( <i>Argulus</i> , Cirripedia))	0.65	0.73
( <i>Argulus</i> , Cirripedia)	1.0	1.0
(Ostracoda, Malacostraca)	-	0.99
((Ostracoda, Malacostraca)(( <i>Derocheilocaris</i> , <i>Raillietiella</i> )( <i>Speleonectes</i> ( <i>Argulus</i> , Cirripedia))))	-	0.61
(Malacostraca( <i>Raillietiella</i> (( <i>Hutchinsoniella</i> , Branchiopoda)(Copepoda, Hexapoda))))	0.44	-
Malacostraca	1.0	1.0
( <i>Raillietiella</i> (( <i>Hutchinsoniella</i> , Branchiopoda)(Copepoda, Hexapoda)))	0.60	-
(( <i>Hutchinsoniella</i> , Branchiopoda)(Copepoda, Hexapoda))	0.65	-
( <i>Hutchinsoniella</i> , Branchiopoda)	0.59	-
Branchiopoda	1.0	1.0
(Copepoda, Hexapoda)	0.67	-
((Copepoda)(( <i>Lepisma</i> , <i>Hutchinsoniella</i> )(remaining hexapod taxa)))	-	0.70
(( <i>Lepisma</i> , <i>Hutchinsoniella</i> )(remaining hexapod taxa))	-	0.58
Hexapoda	0.96	-
Entognatha: ((Protura, Diplura)(Collembola))	0.98	-
Nonoculata: (Protura, Diplura)	0.98	1.0
(( <i>Lepisma</i> , <i>Hutchinsoniella</i> )(Protura, Diplura))	-	0.72
( <i>Lepisma</i> , <i>Hutchinsoniella</i> )	-	0.72
Protura	1.0	1.0
Diplura	1.0	1.0
Collembola	1.0	1.0
Ectognatha: (Archaeognatha(Zygentoma, Pterygota)	1.0	-
(Archaeognatha( <i>Ctenolepisma</i> , Pterygota))	-	1.0
Archaeognatha	1.0	1.0
Zygentoma	0.98	-
Dicondylia: (Zygentoma, Pterygota)	0.99	-
( <i>Ctenolepisma</i> , Pterygota)	-	0.99
Pterygota	0.97	0.94
Chastomyaria: (Ephemeroptera, Neoptera)	0.96	0.97
Neoptera	0.98	1.0
(((( <i>Acheta</i> , Mantophasmatodea)(Phasmatodea)) remaining orthoperans)(Hemiptera))	0.62	0.78
Hemiptera	1.0	1.0
((( <i>Acheta</i> , Mantophasmatodea)(Phasmatodea)) remaining orthoperans)	0.81	0.98
(( <i>Acheta</i> , Mantophasmatodea)(Phasmatodea))	0.82	1.0
( <i>Acheta</i> , Mantophasmatodea)	0.81	0.99
Phasmatodea	1.0	1.0
Mantophasmatodea	1.0	1.0
Orthoptera without <i>Acheta</i>	0.99	1.0
((Dermaptera, Plecoptera)(Dictyoptera))	0.42	-
Dictyoptera	1.0	1.0
((Mantodea( <i>Blattella</i> , <i>Gromphadorhina</i> ))(Ectobius, Isoptera))	1.0	1.0
(Mantodea( <i>Blattella</i> , <i>Gromphadorhina</i> ))	0.53	0.55
( <i>Ectobius</i> , Isoptera)	0.89	0.94
Mantodea	1.0	1.0
Blattaria	-	-
((Dermaptera, Plecoptera)(Dictyoptera))(Holometabola))	0.39	-
((Dermaptera, Plecoptera)(Holometabola))	-	0.38
(Dermaptera, Plecoptera)	1.0	1.0
Holometabola	1.0	1.0
(Hymenoptera, remaining holometabolans)	1.0	1.0
Hymenoptera	0.80	0.90
(Coleoptera(( <i>Merope</i> ( <i>Boreus</i> , Siphonaptera))(Lepidoptera, Trichoptera))(Diptera))	0.65	-
(Coleoptera( <i>Merope</i> ( <i>Boreus</i> , Siphonaptera))(Lepidoptera, Trichoptera)(Diptera))	-	0.76
((Merope( <i>Boreus</i> , Siphonaptera))(Lepidoptera, Trichoptera))(Diptera))	0.76	-
((Merope( <i>Boreus</i> , Siphonaptera))(Lepidoptera, Trichoptera)(Diptera))	-	1.0
((Merope( <i>Boreus</i> , Siphonaptera))(Lepidoptera, Trichoptera))	0.62	-
(Merope( <i>Boreus</i> , Siphonaptera))	0.98	1.0
( <i>Boreus</i> , Siphonaptera)	1.0	1.0
((Lepidoptera, Trichoptera)(Diptera))	-	0.90
(Lepidoptera, Trichoptera)	1.0	1.0

pP: Bayesian posterior probability values

Additional file 4 Detailed flow of the analysis procedure in the software package PHASE-2.0.



## Additional file 5. List of chimeran species for concatenated 18S and 28S rRNA sequences.

in Additional file 1 listed as	28S rRNA				18S rRNA			
	species	subgroup*	family	common name	18S rRNA	subgroup*	family	common name
<i>Dermacentor</i> sp. *	<i>Dermacentor</i> sp.	Ixodoidea	Ixodidae	hardbacked ticks	<i>Dermacentor andersoni</i>	Ixodoidea	Ixodidae	hardbacked ticks
<i>Aphonopelma hentzi</i> *	<i>Aphonopelma hentzi</i>	Mygalomorphae	Theraphosidae	tarantulas	<i>Aphonopelma reversum</i>	Mygalomorphae	Theraphosidae	tarantulas
<i>Artemia</i> sp. *	<i>Artemia</i> sp.	Anostraca	Artemiidae	brine shrimps	<i>Artemia franciscana</i>	Anostraca	Artemiidae	brine shrimps
<i>Bosmina</i> sp. *	<i>Bosmina</i> sp.	Cladocera	Bosminidae	water fleas	<i>Bosmina longirostris</i>	Cladocera	Bosminidae	water fleas
<i>Cyclopidae</i> sp. *	<i>Cyclopidae</i> sp.	Cyclopoida	Cyclopidae	-	<i>Macrocyclus albidus</i>	Cyclopoida	Cyclopidae	-
<i>Penaeus vannamei</i> *	<i>Penaeus vannamei</i>	Dendrobranchiata	Penaeidae	penaeid shrimps	<i>Penaeus semisulcatus</i>	Dendrobranchiata	Penaeidae	penaeid shrimps
<i>Raillietiella</i> sp. *	<i>Raillietiella</i> sp.	Pentastomida	Cephalobaenidae	tongue worms	<i>Raillietiella</i> sp.	Pentastomida	Cephalobaenidae	tongue worms
<i>Baculentulus densus</i> *	<i>Baculentulus densus</i>	Acerentomata	Acerentomidae	-	<i>Baculentulus tianmushanensis</i>	Acerentomata	Acerentomidae	-
<i>Epeorus sylvicola</i> *	<i>Epeorus sylvicola</i>	Setisura	Heptageniidae	flatheaded mayflies	<i>Epeorus longimanus</i>	Setisura	Heptageniidae	flatheaded mayflies
<i>Siphonura aestivalis</i> *	<i>Siphonura aestivalis</i>	Pisciforma	Siphonuridae	fish-bodied mayflies	<i>Siphonura croaticus</i>	Pisciforma	Siphonuridae	fish-bodied mayflies
<i>Mantophasma zephyra</i> *	<i>Mantophasma zephyra</i>	Mantophasmatodea	Mantophasmatidae	heel-walkers	<i>Mantophasma cf. zephyra</i>	Mantophasmatodea	Mantophasmatidae	heel-walkers
<i>Hierodula membranacea</i> *	<i>Hierodula membranacea</i>	Mantodea	Mantidae	praying mantids	<i>Hierodula schultzei</i>	Mantodea	Mantidae	praying mantids
<i>Isoperla</i> sp. *	<i>Isoperla</i> sp.	Perloidea	Perlidae	predatory stoneflies	<i>Isoperla obscura</i>	Perloidea	Perlidae	predatory stoneflies
<i>Nemoura flexuosa</i> *	<i>Nemoura flexuosa</i>	Nemouroidea	Nemouridae	spring stoneflies	<i>Nemoura cinerea</i>	Nemouroidea	Nemouridae	spring stoneflies
<i>Pyrrhocoris apterus</i> *	<i>Pyrrhocoris apterus</i>	Heteroptera	Pyrrhocoridae	stainers	<i>Dysdercus poecilus</i>	Heteroptera	Pyrrhocoridae	stainers
<i>Harpocera thoracica</i> *	<i>Harpocera thoracica</i>	Heteroptera	Miridae	plant bugs	<i>Polymerus castilleja</i>	Heteroptera	Miridae	plant bugs
<i>Cercopis vulnerata</i> *	<i>Cercopis vulnerata</i>	Cercopoidea	Cercopidae	spittlebugs	<i>Mahanarva costaricensis</i>	Cercopoidea	Cercopidae	spittlebugs
<i>Anacridium aegypticum</i> *	<i>Anacridium aegypticum</i>	Caelifera	Acrididae	short-horned grasshoppers	<i>Acrida cinerea</i>	Caelifera	Acrididae	short-horned grasshoppers
<i>Leptophyes punctatissima</i> *	<i>Leptophyes punctatissima</i>	Ensifera	Tettigoniidae	katydids	<i>Microcentrum rhombifolium</i>	Ensifera	Tettigoniidae	katydids
<i>Pholidoptera griseocapta</i> *	<i>Pholidoptera griseocapta</i>	Ensifera	Tettigoniidae	katydids	<i>Tettigonia viridissima</i>	Ensifera	Tettigoniidae	katydids
<i>Nomada</i> sp. *	<i>Nomada</i> sp.	Aculeata	Apidae	cuckoo bees	<i>Apis mellifera</i>	Aculeata	Apidae	honey bee
<i>Scolia</i> sp. *	<i>Scolia</i> sp.	Aculeata	Scoliidae	scoliid wasps	<i>Scolia verticalis</i>	Aculeata	Scoliidae	scoliid wasps
<i>Tenthredinidae</i> sp. *	<i>Tenthredinidae</i> sp.	Tenthredinoidea	Tenthredinidae	common sawflies	<i>Dolerus</i> sp.	Tenthredinoidea	Tenthredinidae	common sawflies
<i>Tenebrio</i> sp. *	<i>Tenebrio</i> sp.	Polyphaga	Tenebrionidae	darkling ground beetles	<i>Tenebrio molitor</i>	Polyphaga	Tenebrionidae	darkling ground beetles
<i>Ctenocephalides felis</i> *	<i>Ctenocephalides felis</i>	Pulicomorpha	Pulicidae	common fleas	<i>Ctenocephalides canis</i>	Pulicomorpha	Pulicidae	common fleas
<i>Pieris napi</i> *	<i>Pieris napi</i>	Glossata	Pieridae	Whites and Yellows	<i>Anthocharis sara</i>	Glossata	Pieridae	Whites and Yellows
<i>Oxyethira rossi</i> *	<i>Oxyethira rossi</i>	Spicipalpia	Hydroptilidae	purse casemaker caddisflies	<i>Oxyethira dualis</i>	Spicipalpia	Hydroptilidae	purse casemaker caddisflies
<i>Trianaodes</i> sp. *	<i>Trianaodes</i> sp.	Integripalpia	Leptoceridae	longhorned casemaker caddisflies	<i>Oecetis avara</i>	Integripalpia	Leptoceridae	longhorned casemaker caddisflies
<i>Milnesium</i> sp. *	<i>Milnesium</i> sp.	Apochela	Milnesiidae	water bears	<i>Milnesium tardigradum</i>	Apochela	Milnesiidae	water bears

\* given subgroups have not necessarily the same hierarchical level



**Additional file 6. Primer list 18S rRNA.**

Primer	Direction	Sequence 5' - 3'	Taxa	References
18SL0001	forward	TACCTGGTTGATCCTGCCAGT	AH, My	Luan et al. 2003
1F	forward	TACCTGGTTGATCCTGCCAGTAG	AH, My	Giribet & Ribera 2000
18S1L	forward	TACCTGGTTGATCCTGCCAGT	AH, My	Luan et al. 2005
18SV0000	forward	TACCTGGTGGATCCTGCCAGTA	AH, My	Chalwatzis et al. 1995
18SL0466	forward	GTTTCGATTCCGGAGAGGGAG	AH, My	Luan et al. 2003
3F	forward	GTTTCGATTCCGGAGAGGGGA	AH, My	Giribet et al. 1996
18SL500	forward	GTTTCGATTCCGGAGAGGGAG	AH, My	Luan et al. 2005
18Sai	forward	CCTGAGAAACGGCTACCACATC	AH, My	Maddison et al. 1999
18SL0922	forward	AATTGGAGTGCCTCAAAGCAGGC	AH, My	Luan et al. 2003
5F	forward	GCGAAAGCATTGCGCAAGAA	AH, My	Giribet et al. 1996
18SL1210	forward	CCTTGAGAAAATTGGAGTGCT	AH, My	Luan et al. 2005
18Sbi rev	forward	TCCGATAACGAACGAGACTC	AH, My	De Salle et al. 1992
18SL1362	forward	CTTAATTGACTCAACACGGG	AH, My	Luan et al. 2003
18S3L	forward	AGGAATTGACGGAAGGGCAC	AH, My	Luan et al. 2005
18A1	forward	CCTAYCTGGTTGATCCTGCCAGT	Cr, Py	Dreyer & Wägele 2001
700 F-MR	forward	GCCGCGGTAATTCCAGC	Cr, Py	Raupach, unpubl.
1000 F	forward	CGATCAGATACCGCCCTAGTTC	Cr, Py	Dreyer & Wägele 2001
1250 FN-MR	forward	GGCCGTTCTTAGTTGGTGGAG	Cr, Py	Raupach, unpubl.
18SR0532	revers	TTGCGCGCTGCTGCCTTCC	AH, My	Luan et al. 2003
5R	revers	CTTGGCAAATGCTTTCCG	AH, My	Giribet et al. 1996
18S1R	revers	TAATATACGCTATTGGAGCTGG	AH, My	Luan et al. 2005
18Sbi.0	revers	TAACCGCAACAACCTTAAAT	AH, My	De Salle et al. 1992
18SR1100	revers	CGACGATCCAAGAATTTAC	AH, My	Luan et al. 2003
18Sbi	revers	GAGTCTCGTTTCGTTATCGGA	AH, My	Maddison et al. 1999, Giribet et al. 1999
18SR1470	revers	TTAGAACTAGGGCGGTATCTG	AH, My	Luan et al. 2005
18SR1524	revers	AGTCTCGTTTCGTTATCGGAAT	AH, My	Luan et al. 2003
9R	revers	GATCCTTCCGCAGGTTACCTAC	AH, My	Giribet et al. 1996
18SR1790	revers	CGTTACCGGAATGAACCAGAC	AH, My	Luan et al. 2005
18SR1900	revers	TAATGATCCTTCTGCAGGTTACCTACG	AH, My	Chalwatzis et al. 1995
18SR2090 or 18S3R	revers	CCTACGGAAACCTTGTACG	AH, My	Luan et al. 2003, Luan et al. 2005
700 R	revers	CGCGGCTGCTGGCACCAGAC	Cr, Py	Wollscheid, Dryer & Englisch, unpubl.
1000 R	revers	GAACTAGGGCGGTATCTGATCG	Cr, Py	Dreyer & Wägele 2001
1155 R	revers	CCGTCAATTCTTTAAGTTTCAG	Cr, Py	Dreyer & Wägele 2001
1500 R	revers	CATCTAGGGCATCACAGACC	Cr, Py	Wollscheid, Dryer & Englisch, unpubl.
1800	revers	GATCCTTCCGCAGGTTACCTACG	Cr, Py	Wollscheid, Dryer & Englisch, unpubl.

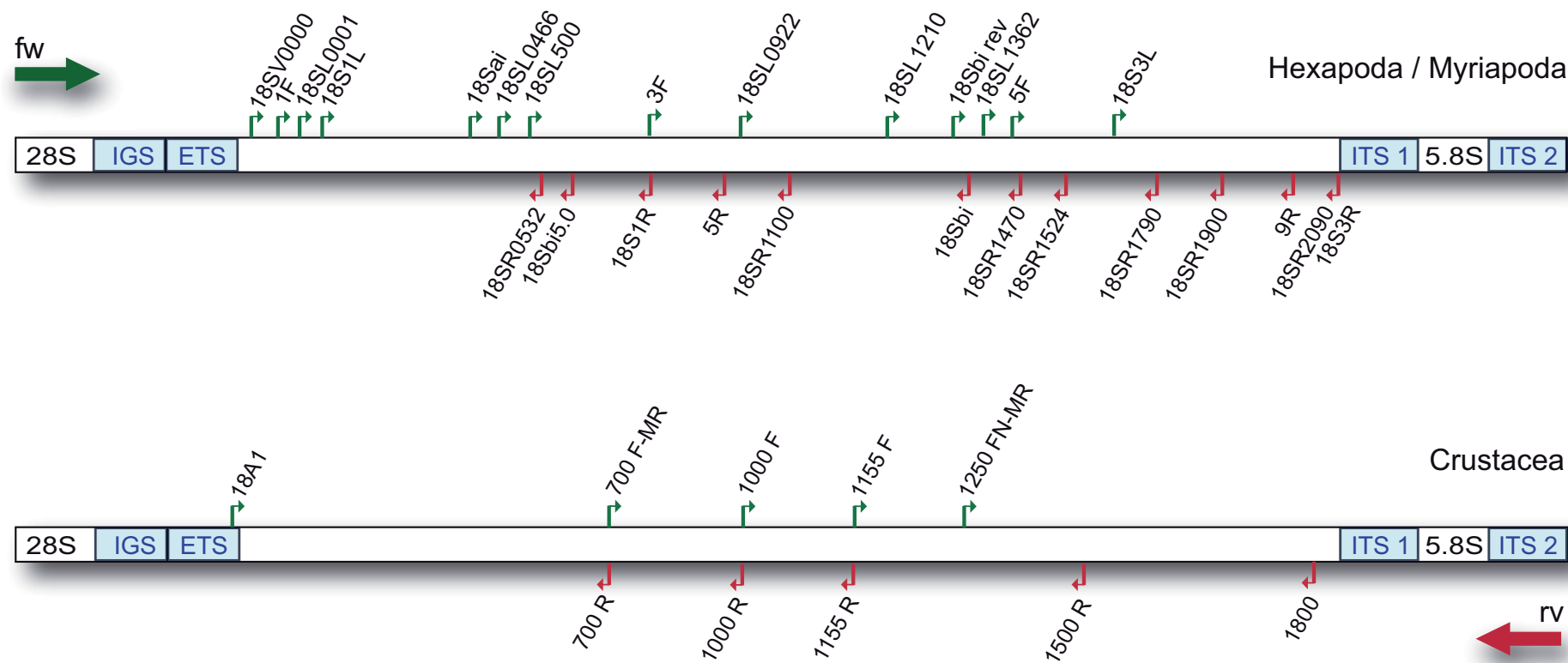
AH: Apterygote hexapods; My: Myriapods; Cr: Crustaceans; Py: Pycnogonids. Description and primer combinations are given in Additional file 8 and 16.

## Additional file 7. Primer list 28S rRNA.

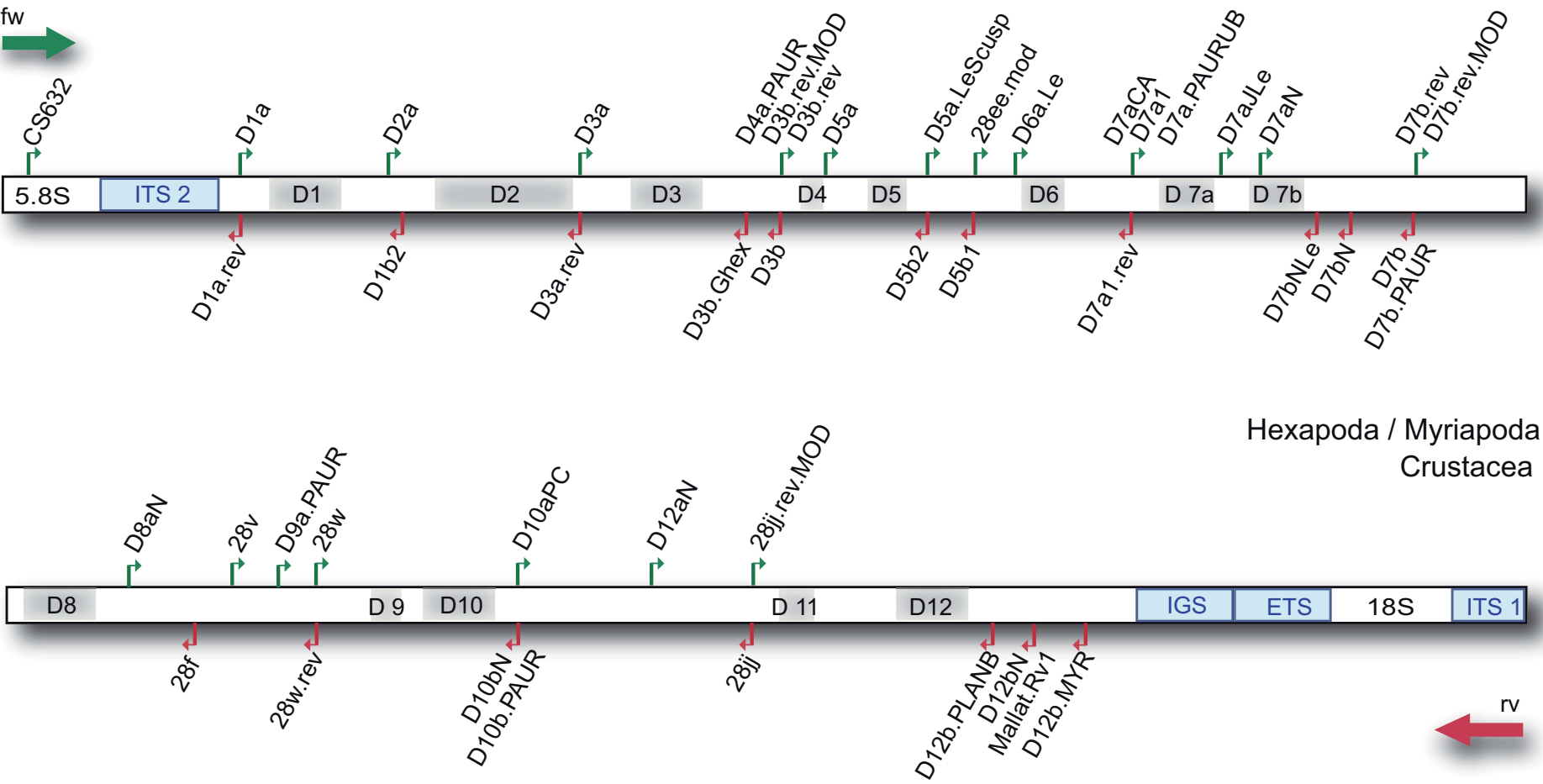
Primer	Direction	Sequence 5' - 3'	Taxa	References
CS632	forward	CGATGAAGAACGCAGC	AH, My, Cr	Schlötterer et al. 1994
427 or D1a	forward	CCC(C/G)CGTAA(T/C)TTAAGCATAT	AH, My, Cr	Friedrich & Tautz 1997
D2a	forward	GATAGCGAACAAAGTACC	AH, My, Cr	Dell'Ampio et al. subm.
D3a	forward	GACCCGTCTTGAAACACGGA	AH, My, Cr	Nunn et al. 1996
D3b.rev.MOD	forward	TAGTAGCTAGTTCCCTCCG	AH, My, Cr	reverse primer of D3b (Nunn et al. 1996), modified, Dell'Ampio et al. subm.
D4a.PAUR	forward	GTTCCCTCCGAAGTTTCC	Pau	Bartel, present study
742 or D5a	forward	CTCAAACCTTTAAATGG	AH, My, Cr	Friedrich and Tautz 1997
D5a.LeScusp	forward	TGGTAAGCAGGACTGG	AH	Dell'Ampio, present study
28ee.mod	forward	CCGTAAGGAGTGTGTAAC	AH, My, Cr	Hillis & Dixon 1991, modified by Dell'Ampio, unpubl. (PHDthesis)
D6a.Le	forward	TGTAACGACTCACCTGC	AH	Dell'Ampio, present study
476 or D7a1	forward	CTGAAGTGGAGAAGGGT	AH, My, Cr	Friedrich and Tautz 1997
D7aCA	forward	CGATGTGGAGAAGGG	AH	Dell'Ampio et al. subm.
D7aN	forward	AGAACCTGGTGACGGAAC	AH, My, Cr	Dell'Ampio, unpubl. (PHDthesis)
D7aJLe	forward	CGAAAGGGGAAGTGGGGTATC	AH	Dell'Ampio, present study
D7a.PAURUB	forward	GCTGAAGTGGAGAAGG	My	Bartel, present study
D7b.rev	forward	ATGTAGGTAAAGGGAAGTC	AH, My, Cr	reverse primer of D7b (Friedrich and Tautz 1997), Dell'Ampio et al. subm.
D7b.rev.MOD	forward	GATCCCTTAACCTCG	AH, My	reverse primer of D7b (Friedrich and Tautz 1997), modified, Dell'Ampio et al. subm.
28v	forward	AAGTAGCCCAAATGCCTCATC	AH, My, Cr	Hillis & Dixon 1991
D8aN	forward	TCAGAACTGGCAGGACCGG	AH	Dell'Ampio, unpubl. (PHDthesis)
D9a.PAUR	forward	AATCAGCGGGGAAAG	Pau	Bartel, present study
28w	forward	CCT(G/T)TTGAGCTTGACTCTAATCTG	AH, My, Cr	Hillis & Dixon 1991
D10aPC	forward	GGGGAGTTTGACTGGGGCGG	AH, My, Cr	Dell'Ampio, present study
28jj.rev.MOD	forward	AGGTTAGTTTTACCTAC	AH, My	reverse primer of 28jj (Hillis & Dixon 1991), modified, Dell'Ampio, present study
D12aN	forward	GAGCAAGAGGTGTGAGAAAAGTTAC	AH, My, Cr	Dell'Ampio, unpubl. (PHDthesis)
28S rD1.2a	forward	CCSSSGTAATTTAAGCATATTAT	Pt	Whiting 2002
28S rD3.2a	forward	AGTACGTGAAACCGTTCCASGGGT	Pt	Whiting 2002
28S rD3.2a.mod	forward	*	Pt	Whiting 2002, modified by Simon, unpubl.
AnsBfor	forward	TCAGAGTCGGGTGCTTGAGG	Od	Kück, unpubl. (Masterthesis)
CB1.2a	forward	AAACTCCACCTAAGACTGAATACGA	Libellulidae	Schmidt, unpubl. (Masterthesis)
PB1.a	forward	TAAACTCCAYCTAAGACTGAC	Aeshnidae	Letsch, unpubl. (PHDthesis)
28S A	forward	GACCCGTCTTGAAGCACGT	Pt	Whiting 2002
28S A.mod	forward	*	Pt	Whiting 2002, modified by Simon, unpubl.
Ans2.1a	forward	TCGTCNNGAGCTGGGTATG	Od	Letsch, unpubl. (PHDthesis)
Ans3.1a	forward	DHAANGGGTTCGTACAGT	Od	Letsch, unpubl. (PHDthesis)
Ans4.1a	forward	CGGCTACCTTAAGAGAGTC	Od	Letsch, unpubl. (PHDthesis)
N4upu	forward	KTGCCAGGTRSGGAGTTTG	Od	Letsch, unpubl. (PHDthesis)
28S Rd4.5a	forward	AAGTTTCCCTCAGGATAGCTG	Od	Whiting 2002
28ee	forward	ATCCGCTAAGGAGTGTGTAACTCACC	Od	Hillis & Dixon 1991
28ll	forward	*	Pt	Hillis & Dixon 1991, modified by Simon, unpubl.
28w	forward	*	Pt	Hillis & Dixon 1991, modified by Simon, unpubl.
28y	forward	*	Pt	Simon, unpubl.
D1a.rev	revers	ATATGCTTAAATTAAGCGGG	AH, My, Cr	reverse primer of D1a (Friedrich and Tautz 1997), Dell'Ampio, present study
D1b2	revers	CGTACTATTGAACCTCTCTCTT	AH, My, Cr	Dell'Ampio et al. 2002
D3a.rev	revers	TCCGTGTTTCAAGACGGGAC	AH, My, Cr	reverse primer of D3a (Nunn et al. 1996), Dell'Ampio, unpubl. (PHDthesis)
D3b	revers	TCCGGAAGGAACCAAGCTACTA	AH, My, Cr	Nunn et al. 1996
D3b.Ghex	revers	GAATCGCTAAGGACCTCC	<i>G. hexasticha</i> , Pau	Bartel, present study
706 or D5b2	revers	CGCCAGTCTCGCTTACC	AH, My, Cr	Friedrich and Tautz 1997
689 or D5b1	revers	ACACACTCCTTAGCGGA	AH, My, Cr	Friedrich and Tautz 1997
D7a1.rev	revers	AAACCCCTTCTCCACATCGG	AH, My, Cr	reverse primer of D7a1.rev (Friedrich & Tautz 1997), Dell'Ampio et al. subm.
477 or D7b	revers	GACTTCCCTTACCTACAT	AH, My, Cr	Friedrich and Tautz 1997
D7bNLe	revers	GGACCCGACGATTCTC	Cr, <i>L. saccharina</i>	Dell'Ampio, present study
D7b.PAUR	revers	ATCCTTTTCGCCGAAG	Pau	Bartel, present study
23 or 28f	revers	CAGAGCACTGGGCGAGAAATCAC	AH, My, Cr	Van der Auwera et al. 1994, modified by Dell'Ampio, unpubl. (PHDthesis)
28w.rev	revers	CAGATTAGAGTCAAGCTCAACAGG	AH, My, Cr	reverse primer of 28w (Hillis & Dixon 1991), Dell'Ampio et al. subm.
28jj	revers	AGTAGGGTAAACTAACCT	AH, My, Cr, Od, Pt	Hillis & Dixon 1991
D10bN	revers	TTTGACAGATGTACCCGCC	Cr, Neanuridae	Dell'Ampio, unpubl. (PHDthesis)
D10b.PAUR	revers	ACCATTTGACAGATGTACCGCC	Pau	Bartel, present study
D12b.PLANB	revers	GAGTACGACACCCC	AH, My, Cr	Dell'Ampio, present study
D12bN	revers	TATGGCAGCTGCTCTACC	AH, My, Cr	Dell'Ampio, unpubl. (PHDthesis)
Mallat.Rv1	revers	ACTTTCAATAGATCGCAG	AH, My, Cr	Mallat and Sullivan 1998 (cited as „new primer“)
D12b.MYR	revers	GTTGGTGGCTGCTCTAC	My	Dell'Ampio, present study
28hh	revers	*	Pt	Hillis & Dixon 1991, modified by Simon, unpubl.
28mm	revers	*	Pt	Hillis & Dixon 1991, modified by Simon, unpubl.
28S Rd6.2b	revers	AATAKKAACCRGATTCCTTTTCGCA	Pt	Whiting 2002
28S B	revers	TCGGAAGGAACCAAGCTACA	Pt	Whiting 2002
28S B.mod	revers	*	Pt	Whiting 2002, modified by Simon, unpubl.
28S Rd4.2b	revers	CCTTGGTCCGTGTTTCAAGACGG	Pt	Whiting 2002
28z	revers	*	Pt	Simon, unpubl.
AnsBrev	revers	RGYGGCCCTTCACTTCAT	Od	Kück, unpubl. (Masterthesis)
Cbrev	revers	AGGGCGACCTTCACTTTCATTGC	Libellulidae	Schmidt, unpubl. (Masterthesis)
PB2.b	revers	YACTTTTCATYKTYGCCTATGK	Aeshnidae	Letsch, unpubl. (PHDthesis)
Ans2.2b	revers	GCTCATGCGNAGAAAAGAACTCTA	Od	Letsch, unpubl. (PHDthesis)
Ans3.2b	revers	ATGCTTTGTTTTAATTAGACAGTCA	Od	Letsch, unpubl. (PHDthesis)
Ans4.2b	revers	AGGNAAGAGCCGACATCGAAGGATA	Od	Letsch, unpubl. (PHDthesis)
N4low	revers	TAGAGCGCTTCAGGCATAATC	Od	Letsch, unpubl. (doc thesis)

AH: Apterygote hexapods; My: Myriapods; Pau: Pauropodidae sp.; Cr: Crustaceans; Od: Odonates; Pt: Pterygote insects. Description and primer combinations are given in Additional file 9 and 10.  
 \* sequences upon request from S. Simon, ITZ Hannover, Germany, Email: sabrina.simon@ecolevol.de

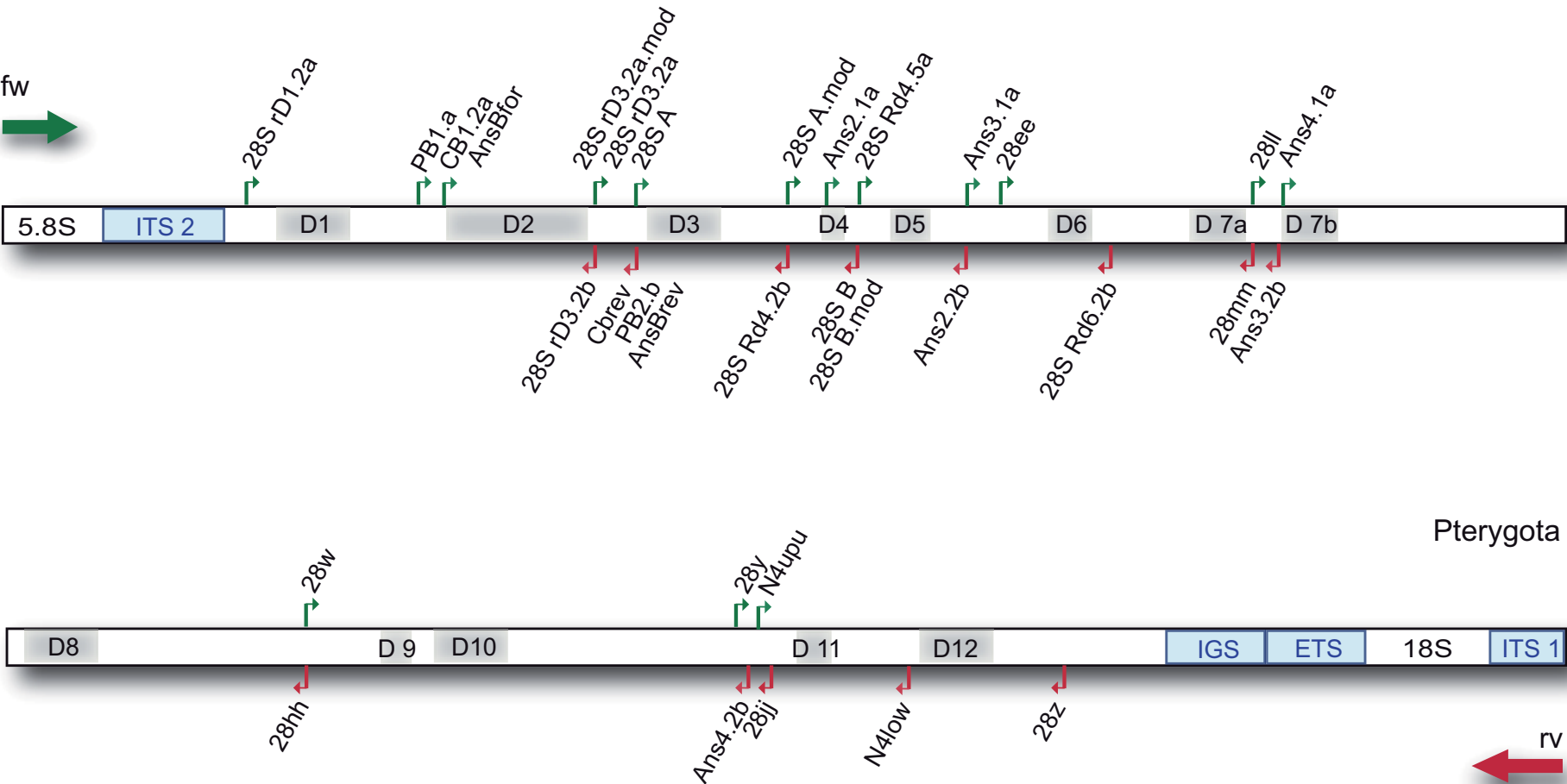
Additional file 8 Primercard of the 18S rRNA gene for hexapods, myriapods and crustaceans.



Additional file 9 Primercard of the 28S rRNA gene for crustaceans, hexapods and myriapods.



Additional file 10 Primercard of the 28S rRNA gene for pterygots.



**Additional file 11**

Supplementary information for lab work (amplification, purification and sequencing of PCR products)

**Amplification of PCR products**

Unless otherwise noted all applied protocols refer to manufacturers advices. All used primers are given in Additional file 9 and 10 [1-18]. Splitting the 18S rRNA of apterygote hexapods and myriapods in three or four fragments following primer settings were used: A) 1F/5R, 3F/18Sbi, 5F/9R B) 18S L 0001/18S R 0532, 18S L 0466/18S R 1100, 18S L 0922/18S R 1524, 18S L 1362/18S R 2090 C) 18S 1 L/18S 1 R, 18S L 500/18S R 1470, 18S L 1210/18S R 1790, 18S 3 L/18S 3 R D) 18SV0000/18Sbi5.0, 18Sai/18Sbi or alternative 18Sai/18SR1900, 18Sbi rev/18SR1900. We combined primer pairs of primer sets A, B, C and D to amplify the complete 18S. The 18S in crustaceans was amplified in one PCR product (18A1/1800) and sequenced with eight primers (700 F, 1000 F, 1155 F, 1250 FN, 700 R, 1000 R, 1155 R and 1500 R), see Additional file 8. The PCR-Multiplex-Kit (Qiagen) was used to prevent pooling of weak PCR products. Hotstart PCRs [19] for 18S apterygote hexapods and myriapods were more success fully (Additional file 12). The nuclear 28S rRNA gene of crustaceans, apterygote hexapods and myriapods was amplified in nine overlapping fragments using following primer combinations: CS632/D1b2, D1a/D3b, D2a/D3a.rev, D3a/D5b1, D5a/D7b, D7a1/28f, D7b.rev/28w.rev, 28v - 28jj and D12aN/D12bN or alternatively D12aN/D12bPLANB, D12aN/D12bMYR or D12aN/MallatRv1. Different primer combinations were used when ever necessary for specific taxa. This was essential for the divergent domain D12. Alternative combinations for crustaceans are: D3b.rev/D5b2, D3b.rev/D5b1, D1a/D5b1, D1a/D5b2, D2a/D5b1, D2a/D5b2, D3a/D5b2, D7a1/28f, D7b.rev/28f, D7brev/D10bN, D10aPC/D12bN and D12aN/D12b.PLANB. To complete the 28S for *Eosentomon sakura* following primer combinations were necessary: D1a/D1b2, D2a/D3a.rev, D2a/D3b and 28ee.mod/D7b (Additional file 9). Based on two sets of universal primers [10, 12], specific primers were designed to amplify 28S sequences of Odonata in seven overlapping fragments. The conserved part of the 5'-end was covered by the universal primer sets Rd1.2a/rD3.2a and 28A/28B (Additional file 10), interrupted by the highly variable regions D2 and D3, for which three specifically adapted primer sets were used: AnsBfor/AnsBrev for standard, CB1.2a/CBrev for Libellulidae and PB1.a/PB2.b for some taxa in Aeshnoidea. The domains II – IV were covered by three specific primer sets: Ans2.1a/Ans2.2b, Ans3.1a/Ans3.2b and Ans4.1a/Ans4.2b, respectively. At the 3'-end the primers N4Lfor/N4Lrev amplified the conserved region before the variable D12 domain (Additional file 7). 28S PCR-products for Pterygota were amplified with following primer combinations: 28S rD1.2a/28S Rd4.2b, 28S rD3.2a/28S B, 28S A/28S Rd6.2b, 28y/28z, 28ll/28hh and 28w/28jj (Additional file 10). The amount of Taq polymerase (Bioline) was increased to 0.3 µl, respectively the amount of sterile H<sub>2</sub>O was diminished for the two last mentioned primer settings (Additional file 13). For the primer pairs 28ee/28hh and 28ll/28jj reagents (Roche) were used with a different composition of PCR- mix (Additional file 13).

**Purification and Sequencing reactions**

Unless otherwise noted all applied protocols refer to manufacturers advices; weak PCR products were pooled for purification. 18SrRNA of apterygotes and myriapods, 28S rRNA of crustaceans and odonates: Products were purified with the NucleoSpin Extract II (Macherey-Nagel) or with enzymes ExoI/SAP. 0.12µl ExoI (20u/µl, Biolabs), 0.45µl SAP (Shrimp Alkaline Phosphatase, 1u/µ, Promega) and 2.43µl RNase-free sterile water was mixed on ice. 3µl of the mixture was added to 10µl of PCR product and incubated for 15min. at 37°C following 20min. 75°C incubation time and cooling down to 12°C. Purified products were checked on agarose gel. To estimate the DNA concentration a mass marker (BioRad) and Nanodrop Spectrophotometer ND-1000 (peqLab) was used. Cycle sequencing reactions of the 28S Odonata were carried out using BigDye ReadyMix (Applied Biosystems). After standard ethanol-precipitation sequencing products were analyzed on an ABI 377 sequencer (Applied Biosystems). Cycle Sequencing reactions of apterygotes, myriapods and crustaceans were performed using DNA Quick Start Mastermix (Beckman Coulter). CS products were ethanol-precipitated or purified with CleanSeq magnetic bead system (Agencourt) followed by sequencing on Beckman Coulter capillary sequencers

CEQTM 8000 and CEQTM 8800. 28S rRNA PCR-products of apterygote hexapods and myriapods were purified loading samples on a 1% agarose gel (TBE buffer1x). Bands were cut and purified using peqGOLD Gel Extraction Kit (peq-Lab Biotechnologie GmbH). Sequencing was carried out on ABI 3130xl Genetic Analyzer. Sequencing of the 28V/D10bPAUR fragment of Pauropodidae sp. required cloning. Purified PCR products were cloned into pCR2.1-TOPO and chemically transformed into TOP10F' competent cells (2µl TOPO Cloning reaction, 25µl component *E. coli* cells). 15µl of transformation product were spread on selective plates (500µl ampicillin (50µg/ml), 40µl (40mg/ml) X-gal, 40µl (100mM) IPTG) and incubate over night. Twelve colonies were picked and after checking five of them were sequenced using vector primer (M13Rv, M13Fw, TOPO TA Cloning Kit, Invitrogen) on an ABI 3130xl Genetic Analyzer.

PCR-products (28S) of Pterygota were precipitated for purification (2µl 4M NHAc, 240µl 98% ethanol, washing: 1ml 70% ethanol, resuspended in 20µl HPLC H<sub>2</sub>O) for the primer combinations 28SrD1.2a/28SRd4.2b, 28SrD3.2a/28SB and 28SA/28SRd6.2b. PCR-Products amplified with 28ee/28hh were purified with MultiScreen PCR Plate (Millipore), purified products were sequenced at Macrogen (Korea). All PCR primers were used for sequencing also, 28mm was used as reverse sequencing primer. Fragments amplified with primer combinations 28ll/28jj, 28ll/28hh and 28w/28jj were cleaned with MultiScreen PCR Plate (Millipore) System. Cycle Sequencing was carried out using DYEnamic ET Dye Terminator Cycle Sequencing Kit (Amersham Bioscience). The sequencing reactions were purified again with the Montage SEQ Kit (Millipore) and sequenced on a MegaBACE1000 system (Amersham Bioscience). 28y/28z fragments were precipitated (see above). Using BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) cycle sequencing products were purified with Sephadex G-50 Superfine (GE Healthcare) and sequenced on a ABI PRISM 310 Genetic Analyzer (Applied Biosystems).

## References

1. Giribet G, Ribera C: **A review of arthropod phylogeny: New data based on ribosomal DNA sequences and direct character optimization.** *Cladistics* 2000, **16**(2):204-231.
2. Chalwatzi N, Baur A, Stetzner E, Kinzelbach R, Zimmermann FK: **Strongly expanded 18S rRNA genes correlated with a peculiar morphology in the insect order of Strepsiptera.** *Zoology (Jena)* 1995, **98**:115-126.
3. Giribet G, Carranza S, Baglioni J, Riutort M, Ribera C: **First molecular evidence for the existence of a Tardigrada + Arthropoda clade.** *Mol Biol Evol* 1996, **13**(1):76-84.
4. Maddison DR, Baker MD, Ober KA: **Phylogeny of carabid beetles as inferred from 18S ribosomal DNA (Coleoptera: Carabidae).** *Syst Entomol* 1999, **24**:103-138.
5. De Salle R, Gatesy J, Wheeler W, Grimaldi D: **DNA sequences from a fossil termite in Oligo-Miocene amber and their phylogenetic implications.** *Science* 1992, **257**(5078):1933-1936.
6. Dreyer H, Wägele JW: **Parasites of crustaceans (Isopoda: Bopyridae) evolved from fish parasites: molecular and morphological evidence.** *Zoology (Jena)* 2001, **103**:157-178.
7. Giribet G, Edgecombe GD, Wheeler WC: **Arthropod phylogeny based on eight molecular loci and morphology.** *Nature* 2001, **413**(6852):157-161.
8. Schlötterer C, Hauser M-T, von Haeseler A, Tautz D: **Comparative evolutionary analysis of rDNA ITS regions in *Drosophila*.** *Mol Biol Evol* 1994, **11**(3):513-522.
9. Nunn GB, Theissen BF, Christensen B, Arctander P: **Simplicity-correlated size growth of the nuclear 28S ribosomal RNA D3 expansion segment in the crustacean order Isopoda.** *J Mol Evol* 1996, **42**(2):211-223.
10. Hillis DM, Dixon MT: **Ribosomal DNA: molecular evolution and phylogenetic inference.** *The Quarterly Review of Biology* 1991, **66**(4):411-453.
11. Dell'Ampio E: **Il gene codificante per l'rRNA 28S: evoluzione della struttura secondaria ed utilità come marcatore filogenetico in alcune specie della famiglia Neanuridae (Hexapoda, Collembola).** University of Siena, Italy Department of Evolutionary Biology; 2003.

12. Whiting MF: **Mecoptera is paraphyletic: multiple genes and phylogeny of Mecoptera and Siphonaptera**. *Zool Sci* 2002, **31**(1):93-104.
13. Kück P: **Phylogenetic Reconstruction of relationships in the suprafamily Aeshnoidea (Anisoptera). Based on the analysis of 28S rRNA and its secondary structure.** Zoologisches Forschungsmuseum A. Koenig, Molecular Lab, University of Bonn, Germany; 2006.
14. Schmidt C: **Phylogeny of Libellulidae (Insecta: Odonata) based on molecular analysis of the complete nuclear 28S rRNA gene including combined nucleotid (DNA)/doubled (RNA) substitution models.** University of Bonn, Germany, Zoologisches Forschungsmuseum A. Koenig, Molecular Lab; 2006.
15. Letsch HO: **Phylogeny of Anisoptera (Insecta: Odonata): Promises and limitations of a new alignment approach.** University of Bonn, Germany, Zoologisches Forschungsmuseum A. Koenig, Molecular Lab; 2007.
16. Dell'Ampio E, Carapelli A, Frati F: **Secondary structure and sequence variation of the 28S rRNA gene in the Neanuridae, and its utility as a phylogenetic marker: Proceedings of the Xth international Colloquium on Apterygota, \ucesk\'e Bud\uejovice 2000: Apterygota at the Beginning of the Third Millennium.** *Pedobiologia (Jena)* 2002, **46**(3-4):274-283.
17. Van der Auwera G, Chapelle S, De Wachter R: **Structure of the large ribosomal subunit RNA of \textit{Phytophthora megasperma}, and phylogeny of the oomycetes.** *FEBS Lett* 1994, **338**(2):133-136.
18. Mallatt J, Sullivan J: **28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes.** *Mol Biol Evol* 1998, **15**(12):1706-1718.
19. Chou Q, Russell M, Birch DE, Raymond J, Bloch W: **Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-NUMBER amplifications.** *Nucleic Acids Res* 1992, **20**(7):1717-1723.



**Additional file 12. PCR temperature-profiles.**

Profile	Taxa	Temperature profile	Number of cycles	Gene	Thermocycler	Remarks / Primer specification
1	AH, My, Cr	94°C 3:00 min 94°C 0:35 min 60°C 0:30 min, TD -1°C to 45°C 72°C 1:30 min 94°C 0:35 min 50°C 0:30 min 72°C 1:30 min 72°C 10:00 min 4°C	15 cycles  25 cycles	18S, 28S	GeneAmp PCR System 2720, GeneAmp PCR System 2700, (Applied Biosystems) T3000 Thermocycler (Biometra)	Depending on fragments and taxa the 1st annealing temperature varied from 60°C-45°C or 55°C-40°C or 50°C-35°C. In each cycle the temperature was decreased by 1°C.
2	AH, My	95°C 15:00 min 94°C 0:35 min 60°C 1:30 min, TD -1°C to 45°C 72°C 1:30 min 94°C 0:35 min 50°C 1:30 min 72°C 1:30 min 72°C 10:00 min 4°C	15 cycles  25 cycles	18S	GeneAmp PCR System 2720, GeneAmp PCR System 2700, (Applied Biosystems) T3000 Thermocycler (Biometra)	
3	AH, My	95°C 5:00 min 95°C 1:00 min 45°C 1:00 min 72° 1:00 min 72°C 10:00 min 4°C	30 cycles	28S	PRIMUS 96 ADVANCED GRADIENT (peqLab)	
4	AH, My	95°C 5:00 min 95°C 0:30 min 45°C 0:30 min 72°C 0:45 min 95°C 1:00 min 56°C 1:00 min 72°C 1:00 min 72°C 10:00 min 4°C	20 cycles  10 cycles	28S	PRIMUS 96 ADVANCED GRADIENT (peqLab)	
5	AH, My	95°C 5:00 min 95°C 1:00 min 56°C 1:00 min 72°C 1:00 min 95°C 0:30 min 56°C 0:30 min 72°C 0:45 min 72°C 10:00 min 4°C	10 cycles  15 cycles	28S	PRIMUS 96 ADVANCED GRADIENT (peqLab)	
6	Pau	94°C 4:00 min 94°C 1:00 min 55°C 1:00 min 72°C 3:15 min 72°C 10:00 min 4°C	30 cycles	28S	PRIMUS 96 ADVANCED GRADIENT (peqLab)	M13Rv/M13Fw PCR after picking clones
7	Od	94°C 4:00 min 94°C 1:00 min 55°C 1:00 min 72°C 3:15 min 72°C 10:00 min 4°C	30 cycles	28S	GeneAmp PCR System 2700 (Applied Biosystems) T-Gradient (Biometra)	fragments A and C
8	Od	94°C 3:00 min 94°C 0:35 min 70°C 0:30 min, TD -1°C to 50°C 72°C 1:30 min 94°C 0:35 min 50°C 0:30 min 72°C 1:30 min 72°C 10:00 min 4°C	20 cycles  25 cycles	28S	GeneAmp PCR System 2700 (Applied Biosystems) T-Gradient (Biometra)	fragments except A and C

### Appendix 3: Major arthropod relationships inferred from rRNA genes

9	Pt	95°C 2:30 min	28S	9700 (Applied Biosystems)	28S rD1.2a / 28S Rd4.2b
		94°C 0:30 min			
		59°C 0:30 min			
		72°C 2:00 min			
		94°C 0:30 min			
		55°C 0:30 min			
		72°C 2:00 min			
		94°C 0:30 min			
		49°C 0:30 min			
		72°C 2:00 min			
		72°C 2:00 min			
		4°C			
10	Pt	95°C 2:30 min	28S	9700 (Applied Biosystems)	28S rD3.2a / 28S B, 28S A / 28S Rd6.2b
		94°C 0:30 min			
		54°C 0:30 min			
		72°C 2:00 min			
		94°C 0:30 min			
		52°C 0:30 min			
		72°C 2:00 min			
		94°C 0:30 min			
		49°C 0:30 min			
		72°C 2:00 min			
		72°C 2:00 min			
		4°C			
11	Pt	95°C 5:00 min	28S	9600 (PerkinElmer)	28ee / 28hh
		94°C 0:30 min			
		62°C 0:30 min			
		72°C 3:00 min			
		94°C 0:30 min			
		58°C 0:30 min			
		72°C 3:00 min			
		94°C 0:30 min			
		55°C 0:30 min			
		72°C 3:00 min			
		72°C 7:00 min			
		4°C			
12	Pt	95°C 5:00 min	28S	5700 (Applied Biosystems)	28ll / 28jj
		94°C 0:30 min			
		45°C 0:30 min			
		72°C 3:00 min			
		72°C 2:00 min			
		4°C			
13	Pt	95°C 2:30 min	28S	9600 (PerkinElmer)	28ll / 28hh
		94°C 0:30 min			
		55°C 0:30 min			
		72°C 3:00 min			
		72°C 2:00 min			
		4°C			
14	Pt	95°C 2:30 min	28S	5700 (Applied Biosystems)	28w / 28jj
		94°C 0:30 min			
		45°C 0:30 min			
		72°C 3:00 min			
		72°C 2:00 min			
		4°C			
15	Pt	95°C 2:30 min	28S	9700 (Applied Biosystems)	28y / 28z
		94°C 0:30 min			
		57°C 0:30 min			
		72°C 1:30 min			
		94°C 0:30 min			
		54°C 0:30 min			
		72°C 1:30 min			
		94°C 0:30 min			
		51°C 0:30 min			
		72°C 1:30 min			
		72°C 2:00 min			
		4°C			

AH: Apterygote hexapods; My: Myriapods; Pau: Pauropodidae sp.; Cr: Crustaceans; Od: Odonates; Pt: Pterygote insects.

°C: temperature in Celsius; X:00: time in minutes; TD: touch down

Additional file 13. PCR chemicals.

Taxa	Chemicals	[Concentration]	Volume	Gene	Specifications
AH, My	Reagents (SIGMA)			18S	Hotstart and evtl. different MgCl <sub>2</sub> -gradients if necessary 5% DMSO used as enhancer, PCR-profile 1
	10 x PCR buffer without MgCl <sub>2</sub>		5.0 µl		
	MgCl <sub>2</sub> (1.5 – 14 µl)	[25 mM]	7.0 µl		
	dNTPs	[2 mM]	4.0 µl		
	Primer forward	[10 pmol/µl]	0.8 µl		
	Primer reverse	[10 pmol/µl]	0.8 µl		
	Taq-Polymerase	[5 u/µl]	0.15 µl		
	HPLC-H <sub>2</sub> O		31.3 µl		
	DNA template		1.0 – 2.0 µl		
	total volume		50 µl		
Cr	Reagents (SIGMA)			18S, 28S	28S Cr: Different MgCl <sub>2</sub> -gradients, PCR-profile 1; 18S Cr: DMSO replaced by sterile water
	10 x PCR buffer without MgCl <sub>2</sub>		5.0 µl		
	MgCl <sub>2</sub>	[25 mM]	5.0 µl		
	dNTPs	[2 mM]	4.0 µl		
	DMSO		2.5 µl		
	Primer forward	[10 pmol/µl]	0.8 µl		
	Primer reverse	[10 pmol/µl]	0.8 µl		
	Taq-Polymerase	[5 u/µl]	0.15 µl		
	HPLC-H <sub>2</sub> O		30.75 µl		
	DNA template		1.0 – 2.0 µl		
AH, My, Cr	Reagents (Qiagen)			18S, 28S	18S AH, My, Cr, 28S Cr: PCR-profile 2
	Multiplex Mastermix (incl. mixture of taq, dNTPs, MgCl <sub>2</sub> , reaction buffer		10.0 µl		
	2 µl Q-solution		2.0 µl		
	1.6 µl Primer forward	[10 pmol/µl]	1.6 µl		
	1.6 µl Primer reverse	[10 pmol/µl]	1.6 µl		
	HPLC-H <sub>2</sub> O		4.3 µl		
	DNA template		0.5 – 1.0 µl		
	total volume		20 µl		
AH, My	MgCl <sub>2</sub> Pääbo buffer		2.0 – 3.0 mM	28S	28S AH, My: PCR-profile 3, 4, 5
	each dNTP (A, C, G, T) (FERMENTAS)		0.25 mM		
	Primer forward		0.8 µl		
	Primer reverse		0.8 µl		
	GoTaq Flexi		0.44 u		
	sterile H <sub>2</sub> O to get final volume				
	DNA template		2.0 µl		
	total volume		25 µl		
Od	Reagents (SIGMA)			28S	28S Od: fragments A and C: PCR-profile 6; remaining fragments: PCR-profile 7
	10 x PCR buffer without MgCl <sub>2</sub>		5.0 µl		
	MgCl <sub>2</sub>	[25 mM]	5.0 µl		
	dNTPs	[2 mM]	4.0 µl		
	DMSO		2.5 µl		
	Primer forward	[10 pmol/µl]	0.8 µl		
	Primer reverse	[10 pmol/µl]	0.8 µl		
	Taq-Polymerase	[5 u/µl]	0.15 µl		
	sterile H <sub>2</sub> O to get final volume				
	DNA template		1.0 µl		
Pt	Reagents (BIOLINE)			28S	28S Pt: Primer combinations 28S rD1.2a/28S Rd4.2b: d PCR-profile 6; 28S rD3.2a/28S B; 28S A/28S Rd6.2b: PCR-profile 9; 28y/28z: PCR-profile 14. Amount of Taq increased to 0.3 µl, respectively amount of sterile H <sub>2</sub> O diminished for primer combination 28l/28hh (PCR-profile 12) and 28w/28jj (PCR-profile 13)
	10 x PCR buffer without MgCl <sub>2</sub>		2.5 µl		
	MgCl <sub>2</sub>	[25 mM]	1.25 µl		
	dNTPs	[25 mM]	2.5 µl		
	Primer forward	[10 pmol/µl]	1.25 µl		
	Primer reverse	[10 pmol/µl]	1.25 µl		
	Taq-Polymerase	[5 u/µl]	0.15 µl		
	sterile H <sub>2</sub> O		14.85 µl		
	DNA template		1.0 µl		
	total volume		25 µl		
Pt	Reagents (Roche)			28S	28SPt: Primer combinations 28ee/28hh: PCR-profile 10; 28ll/28jj: PCR-profile 11
	FastStart PCR Master 2x (incl. Taq, MgCl <sub>2</sub> , reaction buffer and dNTPs)		12.5 µl		
	Primer forward	[10 pmol/µl]	1.25 µl		
	Primer reverse	[10 pmol/µl]	1.25 µl		
	sterile H <sub>2</sub> O		9.0 µl		
	DNA template		1.0 µl		
	total volume		25 µl		
Pau	PCR reaction buffer MgCl <sub>2</sub> Pääbo buffer		2.5 µl	28S	28S Pau: primer combination 28V/D10bPAUR
	dNTPs (FERMENTAS-life sciences)	[2.5 mM of each dNTP]	2.0 µl		
	Primer forward	[10 pmol/µl]	2.0 µl		
	Primer reverse	[10 pmol/µl]	2.0 µl		
	GoTaq Flexi		0.088 µl		
	sterile H <sub>2</sub> O		13.912 µl		
	DNA template		2.0 µl		
	total volume		24.5 µl		
Pau: TOPO Cloning	salt solution	200 mM NaCl, 10mM MgCl <sub>2</sub>	0.5 µl	28S	28S Pau: PCR for cloning approach
	TOPO vector (PCR 2.1-TOPO)		0.5 µl		
	PCR product		2.0 µl		
	total volume		3.0 µl		
Pau: Check of Cloning	PCR reaction buffer MgCl <sub>2</sub> Pääbo buffer		2.5 µl	28S	28S Pau: PCR for cloning approach, primer combination M13Rw/M13Fw
	dNTPs (FERMENTAS-life sciences)	[2.5 mM of each dNTP]	2.5 µl		
	Primer forward	[20 nM]	0.5 µl		
	Primer reverse	[20 nM]	0.5 µl		
	GoTaq Flexi		0.088 µl		
	sterile H <sub>2</sub> O		18.912 µl		
	picked colony				
	total volume		25 µl		

AH: Apterygote hexapods; My: Myriapods; Pau: Pauropodidae sp.; Cr: Crustaceans; Od: Odonates; Pt: Pterygote insects.  
PCR, Polymerase chain reaction; HPLC, High Performance Liquid Chromatography; dNTPs, di-Nucleotidetriphosphate

**Additional file 14.** Setting of exchangeability parameters used pre-runs.

<b>Parameters</b>	<b>time-homogeneous preruns (500,000 and 3,000,000 generations)</b>
Model	MIXED
Tree, proposal priority	1
Model, proposal priority	5
Topology changes, proposal priority	10
Branch lengths, proposal priority	40
Model 1, 3, proposal priority	7
Model 2, 4 proposal priority	8
Average rates, proposal priority	1
Frequencies, proposal priority	2
Rate ratios, proposal priority	1
Gamma parameter, proposal priority	1
random seed	new seed set for each run

**Additional file 15.** Included chains to infer the time-heterogeneous consensus tree.

included time-heterogeneous chains	generations (burn-in excluded)	harmonic mean ln-Likelihood	2ln (B <sub>10</sub> )
1	5 million	78999.3698869758	-
2	5 million	78999.6927688722	0.64576379282516
3	5 million	78999.7099518418	0.6801297320053
4	5 million	79000.0584686094	1.37716326719965
5	5 million	79001.9543472591	5.16892056661891
6	5 million	79002.5669121369	6.39405032221111
7	8 million	79002.8227346211	6.90569529062486
8	8 million	79003.4486271688	8.1574803859985
9	5 million	79003.5539878285	8.36820170542342
10	5 million	79004.027674782	9.31557561241789

**Additional file 16.** Included chains to infer the time-homogeneous consensus tree.

Included time homogeneous chains	generations (burn-in excluded)	harmonic mean ln-Likelihood	2ln (B <sub>10</sub> )
1	8 million	79680.9820195926	-
2	5 million	79683.9097337753	5.8554283654
3	5 million	79685.0871292164	8.2102192476

## Additional file 17. Localities of sampled taxa.

Additional file 17. Localities of sampled taxa.

Order	Taxon	Locality	Collection date	Collector	Remarks
Pycnogonida	<i>Colossendeis</i> sp.	ANDEEP I Expedition, Ant XIX-3, Antarctica	29.01.2002	M. Raupach	
Pycnogonida	<i>Nymphon stroemii</i>	Hinlopen Svalbard, Arctica	23.09.2003	d'Dukekem d'Acor	
Notostraca	<i>Triops cancriformis</i>	Marchauen, Austria	2005	E. Eder	
Diplostraca	<i>Daphnia</i> cf. <i>magna</i>	Bonn, Nord-Rhein-Westfalia, Germany	05.10.2005	B. v. Reumont	
	<i>Bosmina</i> sp.	Tegler See, Berlin, Germany	2005	A. Braband	
Ostracoda	<i>Heterocypris incongruens</i>	Hirschweiher, Röttgen, Nord-Rhein-Westfalia, Germany	2005	B. v. Reumont	
	<i>Pontocypris mytiloides</i>	Wilhelmshaven, Niedersachsen, Germany	2007	B. v. Reumont	
Cirripedia	<i>Semibalanus balanoides</i>	Horumersiel, Niedersachsen, Germany	2007	B. v. Reumont	
	<i>Pollicipes pollicipes</i>	Ferrol supermercado, Galicia, Spain	2006	B. v. Reumont	
Branchiura	<i>Argulus</i> cf. <i>foliaceus</i>	Sweden	2007	D. Walošek	
Mystacocarida	<i>Derocheilocaris typicus</i>	Playa dos ninos, Ferrol, Galicia, Spain	2006	B. v. Reumont	
Copepoda	<i>Tigriopus</i> cf. <i>fulvus</i>	Galicia, Spain	2006	B. v. Reumont	
	<i>Canuella perplexa</i>	Hooksiel, Niedersachsen, Germany	09.05.2006	B. v. Reumont	
Remipedia	<i>Speleonectes tulamsensis</i>	Cenote Eden, Puerto Aventuras, Quintana Roo, Mexico	2006	S. Koenemann	
Leptostraca	<i>Nebalia</i> sp.	Ferrol, Galicia, Spain	2006	B. v. Reumont	
Pentastomida	<i>Raillietiella</i> sp.	Asia, host: <i>Hemidactylus</i> cf. <i>frenatus</i>	2007	B. v. Reumont	
Chilopoda	<i>Craterostigma tasmanianus</i>	Tasmania, Australia			
	<i>Lithobius forficatus</i>	Breitenfurt near Vienna, backyard, Niederösterreich, Austria	28.07.2004	N. Szucsich	
Diplopoda	<i>Polyxenus lagurus</i>	Bonn-Plittersdorf, graveyard, Nord-Rhein-Westfalia, Germany	31.05.2005	B. Huber	
	<i>Monographis</i> sp.	Shanghai, China	2005	Y. Yang	
	<i>Polydesmus complanatus</i>	Breitenfurt near Vienna, Niederösterreich, Austria	November 2006	N. Szucsich	
	<i>Cylindroiulus caeruleocinctus</i>	Breitenfurt near Vienna, urban area, Niederösterreich, Austria	23.10.2004	N. Szucsich	
Paupoda	<i>Paurodidae</i> sp.	Panzergraben, Neusiedl am See, Burgenland, Austria	26.04.2006	D. Bartel, N. Szucsich	
Protura	<i>Acerentomon franzi</i>	Lavanttal, Kärnten, Austria	09.10.2005	M. Walz	
	<i>Baculentulus densus</i>	Shinkoji, Sanato, Uda Nagano, Japan	05.05.2006	R. Machida, M. Fikui	
	<i>Eosentomon</i> sp.	Lavanttal, Kärnten, Austria	09.10.2005	M. Walz	
	<i>Eosentomon sakura</i>	Zhanjiang Guangdong, China	2002	Y. Luan, Y. Yang	
	<i>Sinentomon erythranum</i>	Suzhou Jiangsu and Hangzhou Zhejiang, China	2002 - 2006	Y. Luan, Y. Yang	
Diplura	<i>Campodea augens</i>	Breitenfurt near Vienna, forest, Niederösterreich, Austria	01.08.2004	N. Szucsich	
	<i>Lepidocampa weberi</i>	Shinoda, Shizuoka, Japan	20.03.2006	K. Sekiya, R. Machida	
	<i>Cataglyphis aquilonaris</i>	Leopoldsborg XIX. Bezirk, Vienna, Austria	11.11.2004	M. Hable	
	<i>Parajapyx emeryanus</i>	Shanghai, China	2005	Y. Luan, Y. Yang	
	<i>Octostigma sinensis</i>	Zhanjiang Guangdong, China	2002	Y. Yang	
Collembola	<i>Tetradontophora bielanensis</i>	Görlitz, Sachsen, Germany	2006	W. Dunger	
	<i>Gomphiocephalus hodgsoni</i>	Victoria Land, Antarctica		F. Frati	
	<i>Billobella aurantiaca</i>	Feniglia, Grosseto, Toscana, Italy	2000	E. Dell'Amplio	
	<i>Anurida maritima</i>	Livorno, Toscana, Italy		R. Dallai	28S
	<i>Anurida maritima</i>	Texel, ferryport, Noord-Nederland, Netherlands	30.08.2006	K. Meusemann	18S
	<i>Podura aquatica</i>	XXII. Bezirk, Vienna, Austria	27.08.2004	M. Szatecsny, N. Szucsich	28S
	<i>Podura aquatica</i>	T. Hoornje South Texel, Noord-Nederland, Netherlands	30.08.2006	M. Berg	18S
	<i>Cryptopygus antarcticus</i>	Killingbeck Island, Antarctica		A. Carapelli	28S
	<i>Cryptopygus antarcticus</i>	King Georg Islands, Antarctica	2005	M. Raupach	18S
	<i>Isotoma viridis</i>	Rheinbach, Nord-Rhein-Westfalia, Germany	13./14.02.2006	H. Kliebhan	
	<i>Orchesella villosa</i>	Montaluccio, Toscana, Italy	15.09.2004	E. Dell'Amplio	
	<i>Pogonognathellus flavescens</i>	Breitenfurt near Vienna, Niederösterreich, Austria	01.08.2004	N. Szucsich	
	<i>Megalothorax minimus</i>	Vienna, Austria	27.04.2004, 18.05.2005	N. Szucsich	
	<i>Sminthurus viridis</i>	Breitenfurt near Vienna, Niederösterreich, Austria	05.08.2004	N. Szucsich	
	<i>Allacma fusca</i>	Feniglia, Toscana, Italy	Autumn 2005	P. P. Fanciulli	
	<i>Dicyrtoma saundersi</i>	Siena, Toscana, Italy		P. P. Fanciulli	
Archaeognatha	<i>Machilis hrabei</i>	Leopoldsborg XIX. Bezirk, Vienna, Austria	02.09.2005	N. Szucsich	
	<i>Lepismachilis y-signata</i>	XIII. Bezirk, Vienna, Austria	14.08.2004	N. Szucsich	
	<i>Pedetontus okajimae</i>	Shimoda, Shizuoka, Japan	20.03.2006	R. Machida	
Zygentoma	<i>Lepisma saccharina</i>	VIII. Bezirk, Vienna, Austria	24.10.2004	W. Moser	28S
	<i>Lepisma saccharina</i>	Burscheid, Nord-Rhein-Westfalia, Germany	01.11.2005	J. Dambach	18S
	<i>Ctenolepisma longicauda</i>	Esprito Santo, Brazil			
Odonata	<i>Brachytron pratense</i>	France			
	<i>Aeshna juncea</i>	France			
	<i>Oxygastra curtisi</i>	France			
	<i>Cordulia aenea</i>	Japan			
	<i>Somatochlora flavomaculata</i>	France			
	<i>Epiophlebia superstes</i>	Japan			
	<i>Progomphus obscurus</i>	USA			
	<i>Sympetrum danae</i>	France			
	<i>Lestes viridis</i>	Germany			
Ephemeroptera	<i>Epeorus sylvicola</i>	Natural History Museum Prague, Czechia	June 2005		
	<i>Siphonura aestivalis</i>	Natural History Museum Prague, Czechia	August 2005		
Phasmatodea	<i>Carausius morosus</i>	breed, India	2004	A. Melber	
	<i>Bacillus rossius</i>	Tunisia	April 2006	S. Sagasser	
Mantophasmatodea	<i>Mantophasma zephyra</i>	breed, South Africa	2005	R. Predel	
	<i>Tyrannophasma gladiator</i>	breed, South Africa	2006	R. Predel	
Mantodea	<i>Hierodula membranacea</i>	breed, Germany	2006		
Blattaria	<i>Ectobius lapponicus</i>	Hannover Niedersachsen Germany	June 2006	A. Melber	
Dermaptera	<i>Forficula auricularia</i>	Hannover, Niedersachsen, Germany	July 2006	A. Melber	
Plecoptera	<i>Isoperla</i> sp.	Natural History Museum Prague, Czechia	July 2005		
	<i>Nemoura flexuosa</i>	Natural History Museum Prague, Czechia	August 2005		
Hemiptera	<i>Pyrrhocoris apterus</i>	Hannover, Niedersachsen, Germany	June 2006	A. Melber	
	<i>Rhaphigaster nebulosa</i>	Hannover, Niedersachsen, Germany	November 2006	A. Melber	
	<i>Harocera thoracica</i>	Hannover, Niedersachsen, Germany	April 2006	A. Melber	
Hemiptera	<i>Cercopis vulnerata</i>	Hannover Niedersachsen Germany	June 2006	A. Melber	
Orthoptera	<i>Anacridium aegyptium</i>	Tunisia	April 2006	S. Sagasser	
	<i>Leptophyes punctatissima</i>	Hannover, Niedersachsen, Germany	June 2006	A. Melber	
	<i>Pholidoptera griseoptera</i>	Hannover, Niedersachsen, Germany	July 2006	A. Melber	
Hymenoptera	<i>Nomada</i> sp.	Hannover, Niedersachsen, Germany	April 2006	S. Simon	
	<i>Scolia</i> sp.	Tunisia	April 2006	S. Sagasser	
	<i>Tenthredinidae</i> sp.	Hannover, Niedersachsen, Germany	June 2006	A. Melber	
Coleoptera	<i>Silpha obscura</i>	Hannover Niedersachsen Germany	June 2006	A. Melber	
Siphonaptera	<i>Ctenocephalides felis</i>	breed, Germany	2006	C. Epe	
Mecoptera	<i>Boreus hyemalis</i>	Soltau, Niedersachsen, Germany	November 2005	A. Melber	
Lepidoptera	<i>Pieris napi</i>	Hannover, Niedersachsen, Germany	July 2006	A. Melber	
Trichoptera	<i>Trienodes</i> sp.	Hannover, Niedersachsen, Germany	September 2006	S. Simon	

**Additional file 1:** Primer list.

<b>Primer</b>	<b>PrimerSeq (5'-3')</b>
28s-rD1.2a	f- CCC ssG TAA TTT AAG CAT ATT A
28s-rD3.2a	f- AGT ACG TGA AAC CGT TCr sGG GT
28sA	f- GAC CCG TCT TGA ArC ACG
28s-rD4.5a	f- AAG TTT CCC TCA GGA TAG CTG
28ee	f- ATC CGC TAA GGA GTG TGT AAC AAC TCA CC
28II	f- GAT CCG TAA CTT CGG GAY AAG GrT TGG CTC
28v	f- AAG GTA GCC AAA TGC CTC GTC ATC
28w	f- CCT GTT GAG CTT GAC TCT AGT yTG
28y	f- ATCCTTCGATGTCGGCTCTTCC
28jj	r- AGT AGG GTA AAA CTA ACC T
28hh	r- CAr ACT AGA GTC AAG CTC AAC AGG
28gg	r- GAT GAC GAG GCA TTT GGC TAC CTT
28mm	r- GAG CCA AyC CTT rTC CCG AAG TTA CGG ATC
28s-rD6.2b	r- AAT Akk AAC CrG ATT CCC TTT CGC
28sb	r- TCG GAr GGA ACC AGC TAC
28s-rD4.2b	r- CCT TGG TCC GTG TTT CAA GAC GG
28s-rD3.2b	r- yGA ACG GTT TCA CGT ACT mTT GA
28ii	r- GGC TCT TCC TAT CAT TGT GAA GCA GAA TTC AC
28z	r- TGyTCTACCGAGyACAACACC

## Additional file 2: Bayesian support values of 28S rRNA topology for selected clades.

List of Bayesian support values (posterior probability, pP) of the inferred 28S rRNA topology for selected clades of the time-heterogeneous.

Clade	pP: 28S
Entognatha: ((Protura,Diplura)(Collembola))	0.48
Nonoculata: (Protura,Diplura)	1.0
Protura	1.0
Diplura	1.0
Collembola	1.0
Ectognatha: (Archaeognatha(Zygentoma,Pterygota)	0.48
Archaeognatha	1.0
(Pterygota,"Zygentoma")	0.73
(Ephemeroptera,Diptera)	0.92
Ephemeroptera	1.0
Diptera	1.0
(Odonata(Neoptera without Diptera,Zygentoma))	0.52
(( <i>Ctenolepisma</i> ,Dictyoptera)((Phasmatodea(Mantophasmatodea( <i>Acheta</i> , <i>Lepisma</i> ))))(Embioptera,Saltatoria)	0.52
(Ctenolepisma,Dictyoptera)	0.57
Dictyoptera	1.0
((Phasmatodea(Mantophasmatodea( <i>Acheta</i> , <i>Lepisma</i> ))))(Embioptera,Saltatoria)	0.26
(Phasmatodea(Mantophasmatodea( <i>Acheta</i> , <i>Lepisma</i> )))	0.63
(Mantophasmatodea( <i>Acheta</i> , <i>Lepisma</i> ))	0.55
Mantophasmatodea	1.0
(Embioptera,Saltatoria)	0.50
Saltatoria	0.40
(Hemiptera,remaining neopterans)	0.49
Hemiptera	1.0
((Grylloblattodea(Dermaptera,Plecoptera))(Holometabola))	0.41
(Grylloblattodea(Dermaptera,Plecoptera))	0.61
(Dermaptera,Plecoptera)	0.38
Plecoptera	1.0
Holometabola	0.97
(Hymenoptera(Neuropterida,Coleoptera))	0.45
(Neuropterida,Coleoptera)	0.99
Neuropterida	0.75
Coleoptera	1.0
((Siphonaptera,Mecoptera)(Amphiesmenoptera))	0.91
Mecoptera	0.79
Amphiesmenoptera: (Trichoptera,Lepidoptera)	1.0
Trichoptera	1.0
Lepidoptera	1.0

pP: Bayesian posterior probability values



**Additional file 3: Bayesian support values of 18S+28S rRNA topology for selected clades.**

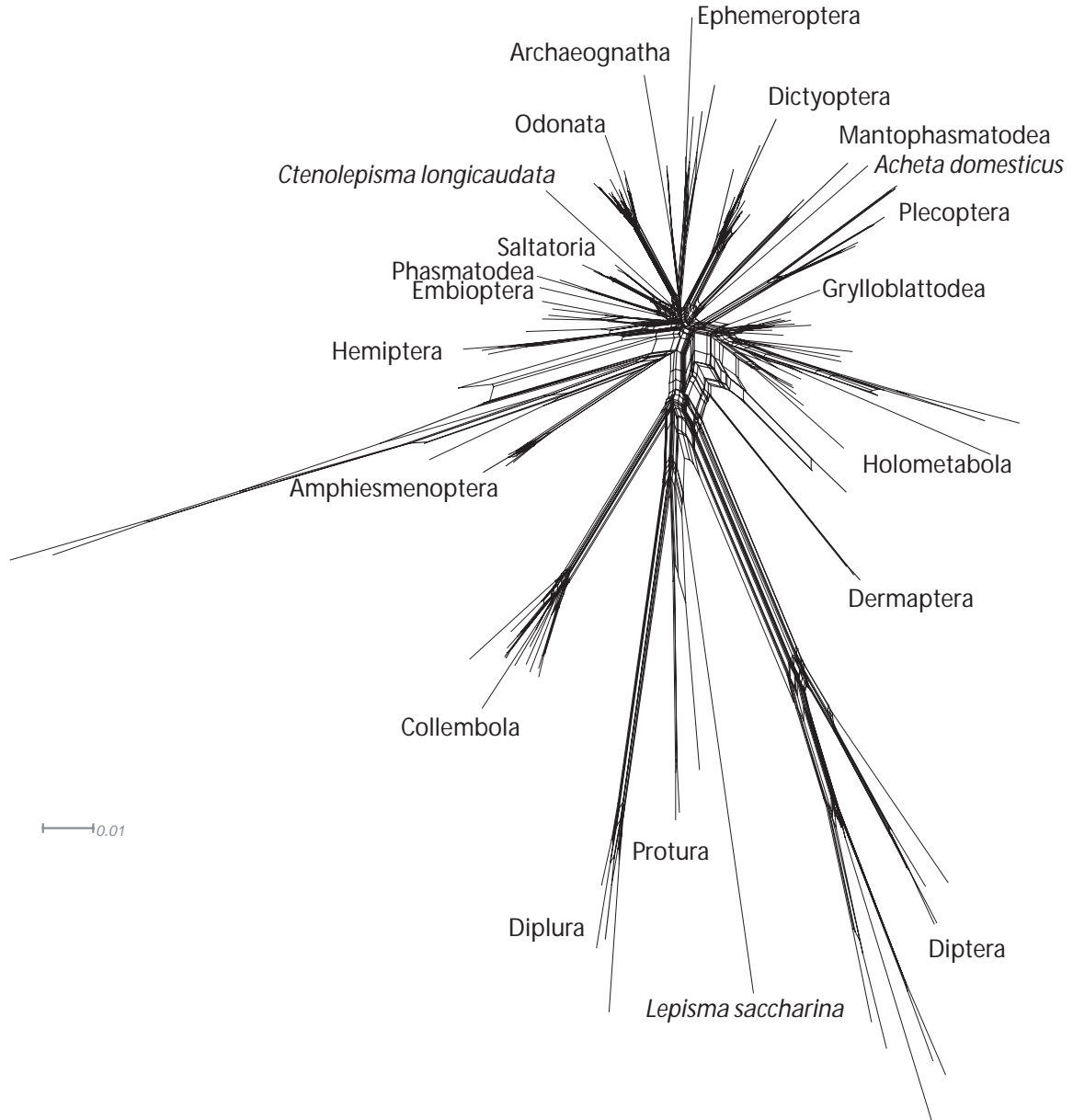
List of Bayesian support values (posterior probability, pP) of the inferred 18S+28S rRNA topology for selected clades of the time-heterogeneous.

Clade	pP: 18S+28S
Entognatha: ((Protura,Diplura)(Collembola))	0.44
Nonoculata: (Protura,Diplura)	1.0
Protura	1.0
Diplura	1.0
Collembola	1.0
Ectognatha: (Archaeognatha(Zygentoma,Pterygota)	0.44
Archaeognatha	1.0
Zygentoma	0.75
Dicondylia: (Zygentoma,Pterygota)	0.87
Pterygota	0.87
Chiasmomyaria: (Ephemeroptera,Neoptera)	0.87
Neoptera	0.87
(Dermaptera,Hemiptera)	0.69
Hemiptera	1.0
(Caelifera,remaining neopterans)	0.71
Caelifera	1.0
(Ensifera without <i>Acheta</i> , remaining neopterans)	0.56
Ensifera without <i>Acheta</i> )	1.0
(Phasmatodea,remaining neopterans)	0.52
Phasmatodea	1.0
( <i>Acheta</i> (Embioptera(((Grylloblattodea,Mantophasmatodea)(Plecoptera))Dictyoptera)))	0.38
(Embioptera(((Grylloblattodea,Mantophasmatodea)(Plecoptera))Dictyoptera))	0.29
((((Grylloblattodea,Mantophasmatodea)(Plecoptera))Dictyoptera)	0.60
((Grylloblattodea,Mantophasmatodea)(Plecoptera))	0.66
(Grylloblattodea,Mantophasmatodea)	1.0
Mantophasmatodea	1.0
Plecoptera	1.0
Dictyoptera	1.0
(Mantodea( <i>Cryptocercus</i> , Isoptera))	0.74
Mantodea	1.0
( <i>Cryptocercus</i> , Isoptera)	1.0
Holometabola	1.0
(Hymenoptera(Neuropterida,Coleoptera))	0.99
Hymenoptera	1.0
(Neuropterida,Coleoptera)	1.0
Neuropterida	1.0
Coleoptera	1.0
(( <i>Boreus</i> (Mecoptera,Siphonaptera))(Diptera,Amphiesmenoptera))	0.99
( <i>Boreus</i> (Mecoptera,Siphonaptera))	0.88
(Mecoptera,Siphonaptera)	0.60
(Diptera,Amphiesmenoptera)	0.66
Amphiesmenoptera: (Trichoptera,Lepidoptera)	1.0
Trichoptera	1.0
Lepidoptera	1.0
Diptera	1.0

pP: Bayesian posterior probability values

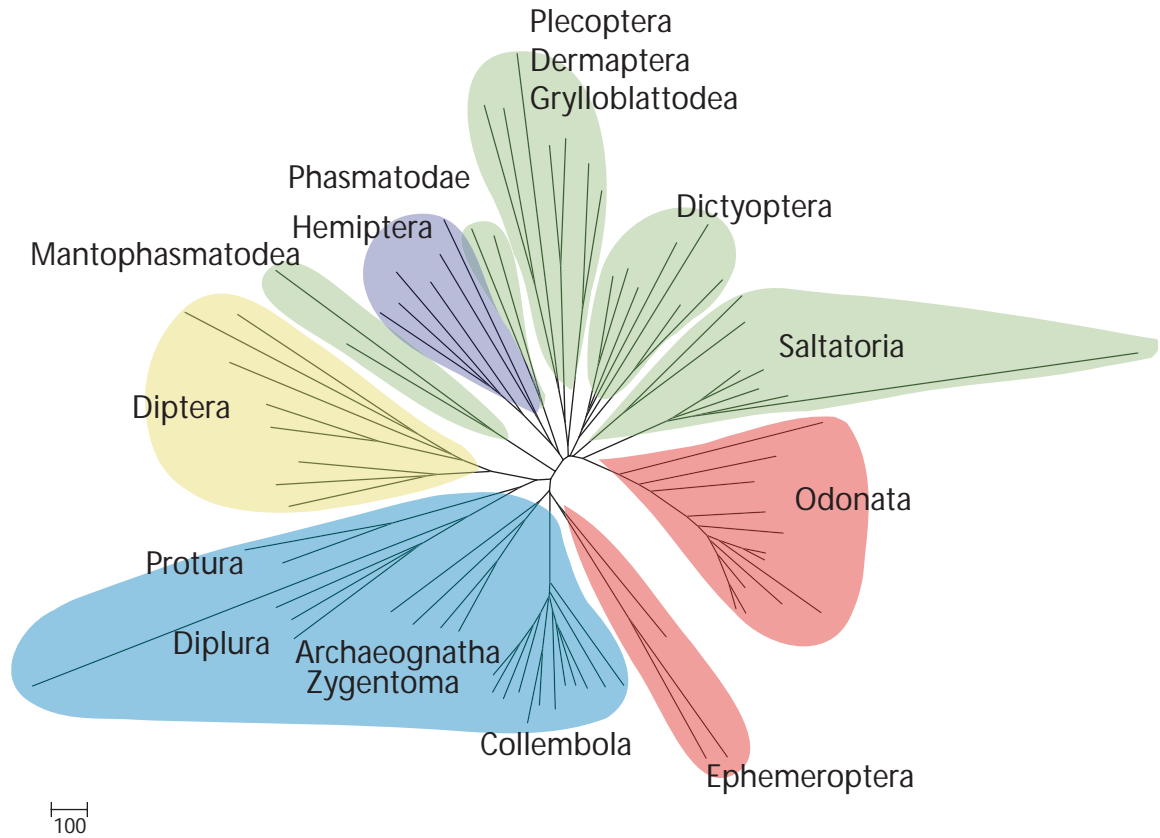
**Additional file 4: Neighbornet graph of the 28S rRNA alignment.**

Neighbornet graph based on uncorrected  $p$ -distances constructed in SplitsTree4 using the 28S rRNA alignment after exclusion of randomly similar sections evaluated with ALISCORE.



**Additional file 5: Geometric distances of the complete 28S secondary structures.**

Complete structures with the exclusion of Mecoptera, Trichoptera, Coleoptera, Lepidoptera and Hymenoptera. Geometric distances of the complete 28S secondary structures, presented by a NJ tree.



**Supplementary table S1** – Taxa list

List of selected taxa for data set creation. Putative orthologs were identified using the new HaMStR approach. \* indicates represented taxa in the final maxspe and maxgen data set respectively.

<b>Taxon</b>	<b>group</b>	<b>Data source</b>	<b>nos. of contigs</b>	<b>nos. of orthologs</b>
<i>Agrotis segetum</i>	Lepidoptera	NCBI	812	158
<i>Anopheles albimanus</i>	Diptera	NCBI	3.096	57
<i>Anopheles anthropophagus</i>	Diptera	NCBI	141	5
<i>Anopheles gambiae</i> *	Diptera	Ensembl	13.133	3096
<i>Antheraea mylitta</i> *	Lepidoptera	NCBI	1.478	193
<i>Aphis gossypii</i> *	Hemiptera	NCBI	3.716	550
<i>Apis mellifera</i> *	Hymenoptera	Ensembl	25.009	3096
<i>Baetis sp.</i> *	Ephemeroptera	this study	3.035	436
<i>Biphyllus lunatus</i>	Coleoptera	NCBI	260	72
<i>Blatella germanica</i>	Dictyoptera	NCBI	1.546	191
<i>Bombyx mori</i> *	Lepidoptera	NCBI	40.444	1490
<i>Danaus plexippus</i> *	Lepidoptera	NCBI	9.930	1178
<i>Diaprepes abbreviatus</i>	Coleoptera	NCBI	1.921	143
<i>Euclidia glyphica</i>	Lepidoptera	NCBI	187	28
<i>Folsomia candida</i>	Collembola	NCBI	5.955	360
<i>Ischnura elegans</i> *	Odonata	this study	3.194	527
<i>Laupala kohalensis</i> *	Ensifera	NCBI	8.371	700
<i>Maconellicoccus hirsutus</i> *	Hemiptera	NCBI	3.929	631
<i>Meladema coriacea</i>	Coleoptera	NCBI	328	60
<i>Melipona quadrifasciata</i>	Hymenoptera	NCBI	321	3
<i>Nasonia giraulti</i> *	Hymenoptera	NCBI	6.764	477
<i>Onychiurus arcticus</i> *	Collembola	NCBI	9.981	755
<i>Plodia interpunctella</i> *	Lepidoptera	NCBI	3.808	431
<i>Plutella xylostella</i> *	Lepidoptera	NCBI	1.048	161
<i>Tenebrio molitor</i>	Coleoptera	NCBI	100	12
<i>Tribolium castaneum</i> *	Coleoptera	NCBI	18.344	1164
<i>Tricholepisma aurea</i>	Thysanura	NCBI	344	85
<i>Vespula squamosa</i>	Hymenoptera	NCBI	1.227	165

### Supplementary table S2 - Genes selected for phylogenetic analysis.

ID – the numerical identifier assigned to the gene during the HaMStR process, FlyBaseID/gene name/symbol – the corresponding ID number/gene name/symbol of the *Drosophila melanogaster* genome database (<http://flybase.org/>). The molecular function and biological process involved description is based on the FlyBase Gene Reports. maxspe/maxgen – genes represented in the alignments. These genes were also selected for the extended ML analyses of individual alignments.

ID number	FlyBaseID	gene name (FlyBase)	symbol	molecular function	biological process involved	maxgen	maxspe
6621	FBgn0002031	lethal (2) 37Cc	Dmel\l(2)37Cc	unknown	unknown	+	
6637	FBgn0014028	Succinate dehydrogenase B	Dmel\SdhB	electron carrier	unknown	+	
6639	FBgn0004907	14-3-3ζ	Dmel\14-3-3ζ	diacylglycerol-activated phospholipid-dependent protein kinase C inhibitor activity	Ras protein signal transduction		+
6671	FBgn0029897	Ribosomal protein L17	Dmel\Rpl17	structural constituent of ribosome	translation	+	+
6692	FBgn0032290	CG6443	Dmel\CG6443	unknown	unknown	+	
6715	FBgn0024558	Diphthamide methyltransferase	Dmel\Dph5	enzyme	unknown	+	
6716	FBgn0024733	Qm	Dmel\Qm	structural constituent of ribosome	translation		+
6754	FBgn0000559	Elongation factor 2b	Dmel\Ef2b	GTPbinding	unknown		+
6790	FBgn0001942	Eukaryotic initiation factor 4a	Dmel\elF-4a	RNA helicase activity	dorsal/ventral axis specification	+	+
6841	FBgn0028336	lethal (1) G0255	Dmel\l(1)G0255	fumarate hydratase activity	tricarboxylic acid cycle	+	
6898	FBgn0033699	Ribosomal protein S11	Dmel\Rps11	structural constituent of ribosome	translation		+
6906	FBgn0034967	eIF-5A	Dmel\elF-5A	translation initiation factor activity	translational initiation	+	+
6910	FBgn0014857	Histone H3.3A	Dmel\His3.3A	DNA binding	cell adhesion	+	+
6913	FBgn0039757	Ribosomal protein S7	Dmel\Rps7	structural constituent of ribosome	translation	+	+
6926	FBgn0035753	Ribosomal protein L18	Dmel\Rpl18	structural constituent of ribosome	translation		+
6927	FBgn0037351	Ribosomal protein L13A	Dmel\Rpl13A	structural constituent of ribosome	translation	+	+
6935	FBgn0010411	Ribosomal protein S18	Dmel\Rps18	structural constituent of ribosome	translation		+
6936	FBgn0038166	CG9588	Dmel\CG9588	protein binding	proteolysis	+	
6951	FBgn0010612	lethal (2) 06225	Dmel\l(2)06225	hydrogen-exporting ATPase activity	proton transport		+
6958	FBgn0036580	PDCD-5	Dmel\PDCD-5	DNA binding	apoptosis	+	+
6972	FBgn0021906	Rieske iron-sulfur protein	Dmel\RFesP	ubiquinol-cytochrome-c reductase activity	mitochondrial electron transport, ubiquinol to cytochrome c	+	+
6978	FBgn0086710	Ribosomal protein L30	Dmel\Rpl30	structural constituent of ribosome	translation		+
6983	FBgn0013954	FK506-binding protein 2	Dmel\FK506-bp2	FK506 binding	protein folding		+
6987	FBgn0034242	CG14480	Dmel\CG14480	unknown	unknown	+	
6990	FBgn0037686	Ribosomal protein L34b	Dmel\Rpl34b	structural constituent of ribosome	translation		+
6999	FBgn0015393	hoi-polloi	Dmel\hoip	mRNA binding	nuclear mRNA splicing, via spliceosome	+	
7007	FBgn0010265	Ribosomal protein S13	Dmel\Rps13	structural constituent of ribosome	translation	+	+
7013	FBgn0020255	ran	Dmel\ran	GTP binding	actin filament organization	+	+
7015	FBgn0086254	CG6084	Dmel\CG6084	aldehyde reductase activity	unknown	+	+
7021	FBgn0052687	CG32687	Dmel\CG32687	protein binding	unknown	+	
7065	FBgn0086656	shrub	Dmel\shrb	unknown	dendrite morphogenesis; protein transport	+	
7083	FBgn0040284	SF2	Dmel\SF2	mRNA binding	nuclear mRNA splicing, via spliceosome	+	
7089	FBgn0002607	Ribosomal protein L19	Dmel\Rpl19	structural constituent of ribosome	translation		+
7098	FBgn0036213	Ribosomal protein L10Ab	Dmel\Rpl10Ab	structural constituent of ribosome	translation	+	+
7171	FBgn0020618	Receptor of activated protein kinase C 1	Dmel\Rack1	protein kinase C binding	oviposition	+	+
7181	FBgn0031148	CG1753	Dmel\CG1753	cystathionine beta-synthase activity	cysteine biosynthetic process via cystathionine	+	
7185	FBgn0039537	CG5590	Dmel\CG5590	oxidoreductase activity	metabolic proces	+	
7188	FBgn0001961	Suppressor of profilin 2	Dmel\Sop2	actin binding	anatomical structure development	+	
7189	FBgn0029176	Ef1γ	Dmel\Ef1γ	translation elongation factor activity	translational elongation	+	+
7214	FBgn0000116	Arginine kinase	Dmel\Argk	arginine kinase activity	phosphorylation	+	+
7236	FBgn0014455	Adenosylhomocysteinase at 13	Dmel\Ahcy13	adenosylhomocysteinase activity	one-carbon compound metabolic process		+

ID number	FlyBaseID	gene name (FlyBase)	symbol	molecular function	biological process involved	maxgen	maxspe
7307	FBgn0023514	CG14805	Dmel\CG14805	unknown	unknown	+	
7310	FBgn0016691	Oligomycin sensitivity-conferring protein	Dmel\Oscp	hydrogen-exporting ATPase activity	proton transport		+
7316	FBgn0034743	Ribosomal protein S16	Dmel\RpS16	structural constituent of ribosome	translation	+	+
7318	FBgn0010078	Ribosomal protein L23	Dmel\RpL23	structural constituent of ribosome	translation		+
7321	FBgn0030968	CG7322	Dmel\CG7322	oxidoreductase activity	metabolic process	+	+
7326	FBgn0076597	-	Dpse\GA16582	structural constituent of ribosome.	translation	+	+
7331	FBgn0047135	CG32276	Dmel\CG32276	unknown	protein modification process	+	+
7339	FBgn0011760	cut up	Dmel\ctp	ATPase activity	microtubule-based movement	+	+
7340	FBgn0052230	CG32230	Dmel\CG32230	NADH dehydrogenase activity	mitochondrial electron transport	+	
7358	FBgn0017579	Ribosomal protein L14	Dmel\RpL14	structural constituent of ribosome	translation		+
7370	FBgn0028342	lethal (1) G0230	Dmel\l(1)G0230	hydrogen-exporting ATPase activity	proton transport		+
7383	FBgn0023211	Elongin C	Dmel\Elongin-C	transcription elongation regulator activity	dendrite morphogenesis	+	
7395	FBgn0000253	Calmodulin	Dmel\Cam	calcium ion binding	kinetochore organization and biogenesis	+	+
7400	FBgn0004432	Cyclophilin 1	Dmel\Cyp1	peptidyl-prolyl cis-trans isomerase activity	protein folding	+	+
7408	FBgn0010408	Ribosomal protein S9	Dmel\RpS9	structural constituent of ribosome	translation		+
7413	FBgn0015790	Rab-protein 11	Dmel\Rab11	GTP binding	anatomical structure development;	+	
7427	FBgn0032597	CG17904	Dmel\CG17904	nucleotide binding	unknown	+	
7430	FBgn0017545	Ribosomal protein S3A	Dmel\RpS3A	structural constituent of ribosome	translation		+
7434	FBgn0039697	CG7834	Dmel\CG7834	electron carrier activity	oxidative phosphorylation	+	+
7437	FBgn0030263	CG2076	Dmel\CG2076	unknown	unknown	+	
7443	FBgn0004117	Tropomyosin 2	Dmel\Tm2	actin binding	heart development		+
7512	FBgn0037346	extra bases	Dmel\exba	protein binding	long-term memory	+	+
7533	FBgn0004867	string of pearls	Dmel\sop	RNA binding	translation	+	+
7538	FBgn0034709	CG3074	Dmel\CG3074	cysteine-type endopeptidase activity	proteolysis	+	+
7542	FBgn0086785	Vps36	Dmel\Vps36	mRNA 3'-UTR binding	unknown	+	
7596	FBgn0019644	ATP synthase, subunit b	Dmel\ATPsyn-b	hydrogen-exporting ATPase activity	proton transport	+	+
7606	FBgn0005593	Ribosomal protein L7	Dmel\RpL7	structural constituent of ribosome	translation	+	+
7609	FBgn0035964	Dihydropteridine reductase	Dmel\Dhpr	6,7-dihydropteridine reductase activity	metabolic process	+	
7631	FBgn0032192	CG5731	Dmel\CG5731	alpha-N-acetylglactosaminidase activity	carbohydrate metabolic process	+	+
7640	FBgn0015282	Proteasome 26S subunit subunit 4 ATPase	Dmel\Pros26.4	ATPase activity	proteolysis	+	+
7660	FBgn0012036	Aldehyde dehydrogenase	Dmel\Aldh	aldehyde dehydrogenase (NAD) activity	pyruvate metabolic process	+	
7720	FBgn0025700	CG5885	Dmel\CG5885	unknown	cotranslational protein targeting	+	+
7731	FBgn0086904	Nascent polypeptide associated complex protein	Dmel\Nacu	protein binding	regulation of pole plasm oskar mRNA localization		+
7736	FBgn0035528	CG15012	Dmel\CG15012	beta-N-acetylhexosaminidase activity.	unknown	+	
7741	FBgn0016119	ATPase coupling factor 6	Dmel\ATPsyn-Cf6	hydrogen-exporting ATPase activity	proton transport		+
7742	FBgn0038739	CG4686	Dmel\CG4686	unknown	unknown	+	
7771	FBgn0023212	Elongin B	Dmel\Elongin-B	transcription elongation regulator activity	protein modification process	+	
7772	FBgn0002579	Ribosomal protein L36	Dmel\RpL36	structural constituent of ribosome	translation		+
7785	FBgn0005533	Ribosomal protein S17	Dmel\RpS17	structural constituent of ribosome	translation		+
7792	FBgn0035871	CG7188	Dmel\CG7188	unknown	negative regulation of apoptosis	+	+
7795	FBgn0000409	Cytochrome c proximal	Dmel\Cyt-c-p	electron carrier activity	oxidative phosphorylation		+
7799	FBgn0002626	Ribosomal protein L32	Dmel\RpL32	structural constituent of ribosome	translation		+
7805	FBgn0037551	CG7891	Dmel\CG7891	GTP binding	small GTPase mediated signal transduction	+	
7864	FBgn0036928	Translocase of outer membrane 20	Dmel\Tom20	P-P-bond-hydrolysis-driven protein transmembrane transporter activity	protein targeting to mitochondrion	+	
7867	FBgn0029785	Ribosomal protein L35	Dmel\RpL35	structural constituent of ribosome	translation		+
7868	FBgn0030733	CG3560	Dmel\CG3560	ubiquinol-cytochrome-c reductase activity	mitochondrial electron transport, ubiquinol to cytochrome c		+
7878	FBgn0031459	CG2862	Dmel\CG2862	nucleotidase activity	unknown		+
7883	FBgn0039713	Ribosomal protein S8	Dmel\RpS8	structural constituent of ribosome	translation	+	+
7884	FBgn0015288	Ribosomal protein L22	Dmel\RpL22	structural constituent of ribosome	translation	+	+
7902	FBgn0037231	CG9779	Dmel\CG9779	unknown	phagocytosis	+	
7903	FBgn0029161	slowmo	Dmel\slmo	unknown	larval behavior	+	
7907	FBgn0035853	CG7375	Dmel\CG7375	ubiquitin-protein ligase activity	regulation of protein metabolic process	+	
7914	FBgn0031090	Rab35	Dmel\Rab35	GTP binding	cytokinesis	+	+
7915	FBgn0036460	CG5114	Dmel\CG5114	unknown	unknown	+	

ID number	FlyBaseID	gene name (FlyBase)	symbol	molecular function	biological process involved	maxgen	maxspe
7932	FBgn0062413	Copper transporter 1A	Dmel\Ctr1A	copper ion transmembrane transporter activity	copper ion transport	+	
7935	FBgn0004404	Ribosomal protein S14b	Dmel\RpS14b	structural constituent of ribosome	translation		+
7950	FBgn0034751	Ribosomal protein S24	Dmel\RpS24	structural constituent of ribosome	translation	+	+
7970	FBgn0001197	Histone H2A variant	Dmel\His2Av	DNA binding	chromatin assembly	+	+
7981	FBgn0035588	CG10672	Dmel\CG10672	oxidoreductase activity	metabolic process	+	
8009	FBgn0035471	Sc2	Dmel\Sc2	oxidoreductase activity	protein modification process	+	+
8013	FBgn0002590	Ribosomal protein S5a	Dmel\RpS5a	structural constituent of ribosome	translation	+	+
8016	FBgn0036318	CG11009	Dmel\CG11009	unknown	unknown	+	
8022	FBgn0015756	Ribosomal protein L9	Dmel\RpL9	structural constituent of ribosome	translation		+
8023	FBgn0010409	Ribosomal protein L18A	Dmel\RpL18A	structural constituent of ribosome	translation	+	+
8030	FBgn0041191	Rheb	Dmel\Rheb	GTP binding	imaginal disc growth	+	
8032	FBgn0010226	Glutathione S transferase S1	Dmel\GstS1	glutathione transferase activity	response to oxidative stress	+	+
8051	FBgn0014026	Ribosomal protein L7A	Dmel\RpL7A	structural constituent of ribosome	translation		+
8073	FBgn0023174	Proteasome $\beta$ 2 subunit	Dmel\Pros $\beta$ 2	endopeptidase activity	ubiquitin-dependent protein catabolic process	+	+
8075	FBgn0003150	Proteasome 29kD subunit	Dmel\Pros29	endopeptidase activity	ATP-dependent proteolysis	+	+
8076	FBgn0033879	CG6543	Dmel\CG6543	enoyl-CoA hydratase activity	fatty acid beta-oxidation	+	+
8090	FBgn0011013	lethal (3) s1921	Dmel\l(3)s1921	deoxyhypusine monooxygenase activity	peptidyl-lysine modification to hypusine	+	
8092	FBgn0030724	Nipsnap	Dmel\Nipsnap	unknown	unknown	+	
8185	FBgn0037001	CG6020	Dmel\CG6020	NADH dehydrogenase (ubiquinone) activity	mitochondrial electron transport, NADH to ubiquinone		+
8207	FBgn0000064	Aldolase	Dmel\Ald	fructose-bisphosphate aldolase activity	glycolysis	+	+
8216	FBgn0033902	Transport and Golgi organization 7	Dmel\Tango7	catalytic activity	Golgi organization and biogenesis		+
8220	FBgn0004169	upheld	Dmel\up	tropomyosin binding	mesoderm development		+
8247	FBgn0001145	Glutamine synthetase 2	Dmel\Gs2	glutamate-ammonia ligase activity	glutamate catabolic process	+	+
8307	FBgn0000579	Enolase	Dmel\Eno	phosphopyruvate hydratase activity	glycolysis	+	
8323	FBgn0024833	AP-47	Dmel\AP-47	protein binding	neurotransmitter secretion	+	
8333	FBgn0015808	Sterol carrier protein X-related thiolase	Dmel\SepX	sterol carrier protein X-related thiolase activity	phospholipid transport	+	
8344	FBgn0031771	CG9140	Dmel\CG9140	NADH dehydrogenase activity	mitochondrial electron transport	+	
8359	FBgn0032444	CG5525	Dmel\CG5525	ATPase activity	mitotic spindle organization and biogenesis	+	
8391	FBgn0011211	bellwether	Dmel\blw	hydrogen-exporting ATPase activity	permatid development		+
8396	FBgn0001098	Glutamate dehydrogenase	Dmel\Gdh	glutamate dehydrogenase [NAD(P)+] activity	sperm storage	+	
8453	FBgn0037893	CG6719	Dmel\CG6719	chaperone binding	de novo' protein folding	+	
8456	FBgn0034138	Ribosomal protein S15	Dmel\RpS15	structural constituent of ribosome	translation		+
8473	FBgn0037874	Translationally controlled tumor protein	Dmel\Tctp	guanyl-nucleotide exchange factor activity	positive regulation of multicellular organism growth	+	+
8474	FBgn0032509	CG6523	Dmel\CG6523	disulfide oxidoreductase activity	cell redox homeostasis	+	
8490	FBgn0033544	CG7220	Dmel\CG7220	ubiquitin-protein ligase activity	proteolysis	+	
8517	FBgn0004926	Eukaryotic initiation factor 2 $\beta$	Dmel\elf-2 $\beta$	translation initiation factor activity	translational initiation		+
8547	FBgn0010638	Sec61 $\beta$	Dmel\Sec61 $\beta$	protein transporter activity.	SRP-dependent cotranslational protein targeting to membrane		+
8565	FBgn0024939	Ribosomal protein L8	Dmel\RpL8	structural constituent of ribosome	translation		+
8581	unknown	unknown	unknown	unknown	unknown	+	+
8607	FBgn0030082	HP1b	Dmel\HP1b	chromatin binding	chromatin assembly	+	+
8609	FBgn0037328	Ribosomal protein L35A	Dmel\RpL35A	structural constituent of ribosome	translation		+
8613	FBgn0019624	Cytochrome c oxidase subunit Va	Dmel\CoVa	cytochrome-c oxidase activity	mitochondrial electron transport, cytochrome c to oxygen		+
8622	FBgn0086687	desat1	Dmel\desat1	stearoyl-CoA 9-desaturase activity	cuticle hydrocarbon biosynthetic process	+	+
8624	FBgn0003279	Ribosomal protein L4	Dmel\RpL4	structural constituent of ribosome	translation		+
8627	FBgn0028690	Rpn5	Dmel\Rpn5	endopeptidase activity	proteolysis	+	+
8640	FBgn0027291	lethal (1) G0156	Dmel\l(1)G0156	socitrate dehydrogenase (NAD+) activity	tricarboxylic acid cycle	+	+
8653	FBgn0022774	Ornithine aminotransferase precursor	Dmel\Oat	ornithine-oxo-acid transaminase activity	ornithine metabolic process	+	+
8657	FBgn0028665	VhaAC39	Dmel\VhaAC39	hydrogen-exporting ATPase activity	proton transport	+	
8661	FBgn0001248	Isocitrate dehydrogenase	Dmel\ldh	isocitrate dehydrogenase (NADP+) activity	glyoxylate cycle	+	
8671	FBgn0033663	ERp60	Dmel\ERp60	protein disulfide isomerase activity	protein folding	+	+
8690	FBgn0086133	knockdown	Dmel\kdn	citrate (Si)-synthase activity	tricarboxylic acid cycle	+	
8714	FBgn0030086	CG7033	Dmel\CG7033	ATP-dependent helicase activity	protein folding	+	
8717	FBgn0022097	Vha36	Dmel\Vha36	hydrogen-exporting ATPase activity	proton transport		+
8732	FBgn0064225	Ribosomal protein L5	Dmel\RpL5	structural constituent of ribosome	translation	+	+

ID number	FlyBaseID	gene name (FlyBase)	symbol	molecular function	biological process involved	maxgen	maxspe
8736	FBgn0023477	Tal	Dmel\Tal	ransaldolase activity	pentose-phosphate shunt	+	+
8740	FBgn0025366	Intronic Protein 259	Dmel\Ip259	unknown	phagocytosis	+	+
8782	FBgn0010602	lesswright	Dmel\lwr	protein binding	regulation of biological process	+	
8784	FBgn0011217	effete	Dmel\eff	protein binding	gamete generation	+	+
8785	FBgn0013325	Ribosomal protein L11	Dmel\RpL11	structural constituent of ribosome	translation		+
8789	FBgn0023175	Proteasome $\alpha$ 7 subunit	Dmel\Prosa7	endopeptidase activity	ubiquitin-dependent protein catabolic process	+	+
8799	FBgn0004922	Ribosomal protein S6	Dmel\RpS6	structural constituent of ribosome	translation	+	+
8804	FBgn0037063	CG9391	Dmel\CG9391	inositol-1(or 4)-monophosphatase activity	dephosphorylation	+	
8822	FBgn0010348	ADP ribosylation factor 79F	Dmel\Arf79F	GTP binding	protein amino acid ADP-ribosylation	+	+
8884	FBgn0014868	Oligosaccharyltransferase 48kD subunit	Dmel\Ost48	dolichyl-diphosphooligosaccharide-protein glycotransferase activity	protein amino acid N-linked glycosylation	+	
8932	FBgn0032987	Ribosomal protein L21	Dmel\RpL21	structural constituent of ribosome	translation		+
8942	FBgn0024188	separation anxiety	Dmel\san	N-acetyltransferase activity	mitotic sister chromatid cohesion; metabolic process	+	
8985	FBgn0025638	Roc1a	Dmel\Roc1a	ubiquitin-protein ligase activity	proteolysis	+	
8986	FBgn0019936	Ribosomal protein S20	Dmel\RpS20	structural constituent of ribosome	translation	+	+
8997	FBgn0039129	Ribosomal protein S19b	Dmel\RpS19b	structural constituent of ribosome	translation	+	+
9007	FBgn0250837	Deoxyuridine triphosphatase	Dmel\dUTPase	dUTP diphosphatase activity	dUTP metabolic process	+	
9017	FBgn0000150	abnormal wing discs	Dmel\awd	microtubule binding	biopolymer modification	+	+
9021	FBgn0039163	CG5515	Dmel\CG5515	unknown	unknown	+	
9093	FBgn0037637	CG9836	Dmel\CG9836	iron-sulfur cluster binding	iron-sulfur cluster assembly	+	
9095	FBgn0028833	Dak1	Dmel\Dak1	cytidylate kinase activity	nucleotide and nucleic acid metabolic process	+	
9097	FBgn0035631	Thioredoxin-like	Dmel\Tx1	disulfide oxidoreductase activity	cell redox homeostasis	+	+
9165	FBgn0029133	REG	Dmel\REG	proteasome activator activity	unknown	+	
9169	FBgn0021814	Vps28	Dmel\Vps28	protein binding	actin cytoskeleton organization and biogenesis	+	
9195	FBgn0014020	Rho1	Dmel\Rho1	GTPase activity; protein binding	anatomical structure development;		+
9284	FBgn0035726	CG9953	Dmel\CG9953	serine-type carboxypeptidase activity	proteolysis	+	
9336	FBgn0003941	Ribosomal protein L40	Dmel\RpL40	structural constituent of ribosome	translation		+
9344	FBgn0037314	Pros $\beta$ 4	Dmel\Pros $\beta$ 4	endopeptidase activity	cell proliferation	+	+
9384	FBgn0011361	mitochondrial acyl carrier protein 1	Dmel\mtacp1	phosphopantetheine binding	mitochondrial electron transport, NADH to ubiquinone	+	
9404	FBgn0031980	Ribosomal protein L36A	Dmel\RpL36A	structural constituent of ribosome	translation		+
9414	FBgn0052672	Autophagy-specific gene 8a	Dmel\Atg8a	unknown	determination of adult life span	+	+
9421	FBgn0032518	Ribosomal protein L24	Dmel\RpL24	structural constituent of ribosome	translation	+	+
9434	FBgn0039132	AP-1 $\sigma$	Dmel\AP-1 $\sigma$	protein transporter activity	neurotransmitter secretion	+	
9438	FBgn0039857	Ribosomal protein L6	Dmel\RpL6	structural constituent of ribosome	translation	+	+
9489	FBgn0002174	lethal (2) tumorous imaginal discs	Dmel\l(2)tid	patched binding	smoothened signaling pathway	+	
9502	FBgn0038742	Arc42	Dmel\Arc42	RNA polymerase II transcription mediator activity	transcription initiation from RNA polymerase II promoter	+	
9503	FBgn0005585	Calreticulin	Dmel\Crc	calcium ion binding	central nervous system development		+
9511	FBgn0014189	Helicase at 25E	Dmel\Hel25E	RNA helicase activity	nuclear mRNA splicing, via spliceosome	+	
9562	FBgn0028985	Serine protease inhibitor 4	Dmel\Spn4	serine-type endopeptidase inhibitor activity	peptide hormone processing	+	+
9569	FBgn0028662	VhaPPA1-1	Dmel\VhaPPA1-1	hydrogen-exporting ATPase activity	mitotic spindle organization and biogenesis	+	
9590	FBgn0028737	Elongation factor 1 $\beta$	Dmel\EFl $\beta$	translation elongation factor activity	translational elongation	+	+
9594	FBgn0250814	-	Dmel\CG4169	ubiquinol-cytochrome-c reductase activity	proteolysis	+	
9603	FBgn0035679	CG10467	Dmel\CG10467	aldose 1-epimerase activity	carbohydrate metabolic process	+	
9616	FBgn0037756	CG8507	Dmel\CG8507	low-density lipoprotein receptor binding	unknown	+	
9666	FBgn0020369	Pros45	Dmel\Pros45	endopeptidase activity	proteolysis	+	
9667	FBgn0036762	CG7430	Dmel\CG7430	dihydrolipoyl dehydrogenase activity	glycine catabolic process	+	+
9684	FBgn0024832	AP-50	Dmel\AP-50	protein binding	neurotransmitter secretion	+	
9751	FBgn0004436	Ubiquitin conjugating enzyme	Dmel\UbcD6	ubiquitin-protein ligase activity	centrosome organization and biogenesis	+	
9753	FBgn0011272	Ribosomal protein L13	Dmel\RpL13	structural constituent of ribosome	translation		+
9813	FBgn0031436	CG3214	Dmel\CG3214	NADH dehydrogenase (ubiquinone) activity	mitochondrial electron transport, NADH to ubiquinone	+	
9821	FBgn0036825	Ribosomal protein L26	Dmel\RpL26	structural constituent of ribosome	translation		+
9826	FBgn0011726	twinstar	Dmel\tsr	actin binding	anatomical structure development	+	+
9827	FBgn0025637	skpA	Dmel\skpA	protein binding	DNA endoreduplication	+	+



**Supplementary table S3** – Maximum likelihood support of individual alignments (assigned with the numerical identifier)

Left: ML support of individual alignments for maxspe data set, right: ML support of individual alignments for maxgen data set. The support for the three different phylogenetic hypotheses of the individual alignments is expressed as the  $\Delta\log L$  : S.E. and the p-SH value. For the best tree the -logL value is given. Tree1 – Palaeoptera hypothesis, Tree2 – Metapterygota hypothesis, Tree3 – Chistostomaria hypothesis.

<i>maxspe</i>				<i>maxgen</i>			
6671	Tree	$\Delta\log L$ :S.E.	p-SH	6637	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-1942,68	1		1	0,0774918	0,592
	2	0,7281781	0,238		2	0,7110949	0,4
	3	0,60541	0,254		3	-2194,7	1
6790	Tree	$\Delta\log L$ :S.E.	p-SH	6671	Tree	$\Delta\log L$ :S.E.	p-SH
	1	1,2008795	0,136		1	-1278,93	1
	2	0,5165057	0,318		2	0,729535	0,215
	3	-3621,21	1		3	0,7295023	0,215
6906	Tree	$\Delta\log L$ :S.E.	p-SH	6715	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,3990349	0,49		1	1,1579939	0,128
	2	-1306,3	1		2	-1619,02	1
	3	0,1648391	0,525		3	1,2706256	0,109
6927	Tree	$\Delta\log L$ :S.E.	p-SH	6790	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,4661826	0,398		1	0,9749671	0,239
	2	-2902,9	1		2	0,2495979	0,502
	3	0,3212636	0,428		3	-2572,28	1
6958	Tree	$\Delta\log L$ :S.E.	p-SH	6906	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,1717124	0,543		1	0,5280528	0,285
	2	0,8339998	0,329		2	0,5280528	0,285
	3	-1719,68	1		3	-982,28	1
7007	Tree	$\Delta\log L$ :S.E.	p-SH	6927	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,6140765	0,437		1	0,4997795	0,435
	2	-1287,81	1		2	-1911,92	1
	3	0,0239833	0,637		3	0,2053388	0,5
7015	Tree	$\Delta\log L$ :S.E.	p-SH	6936	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-3778,23	1		1	-2854,79	1
	2	0,2629581	0,476		2	0,3460208	0,344
	3	0,7776478	0,293		3	0,3460651	0,344
7098	Tree	$\Delta\log L$ :S.E.	p-SH	6958	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-1866,4	1		1	5350	0,249
	2	0,9396534	0,175		2	1,2142472	0,135
	3	0,9395973	0,175		3	-1284,5	1
7214	Tree	$\Delta\log L$ :S.E.	p-SH	7007	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-2653,18	1		1	0,9007565	0,259
	2	1,2555232	0,105		2	0,4062096	0,415
	3	1,2554556	0,105		3	-894,09	1
7316	Tree	$\Delta\log L$ :S.E.	p-SH	7015	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-1252,16	1		1	-2919,77	1
	2	0	0,508		2	0,3881596	0,395
	3	0	0,597		3	0,7457908	0,264
7339	Tree	$\Delta\log L$ :S.E.	p-SH	7098	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-428,72	1		1	-1370,06	1
	2	0,6739587	0,216		2	0,5340454	0,246
	3	0,6739214	0,216		3	0,534081	0,246
7434	Tree	$\Delta\log L$ :S.E.	p-SH	7214	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,7985079	0,223		1	0,0753012	0,487
	2	-2309,35	1		2	0,0750188	0,49
	3	0,6368203	0,237		3	-2321,86	1

<i>maxspe</i>			
7512	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-4812,26	1
	2	1,0238009	0,149
	3	0,0384337	0,136
7538	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,676819	0,26
	2	0,6767644	0,26
	3	-6151,21	1
7606	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,980782	0,3
	2	0,0836479	0,598
	3	-3203,78	1
7631	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-4762,97	1
	2	0,8815169	0,215
	3	0,4547602	0,347
7640	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0	0,113
	2	-1897,11	1
	3	0	0,254
7720	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,7474361	0,264
	2	0,5009634	0,33
	3	-2070,27	1
7883	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,3578732	0,321
	2	0,0028226	0,321
	3	-2093,84	1
7950	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,6400539	0,261
	2	0,0138931	0,276
	3	-1203,47	1
7970	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-489,5	1
	2	0	0,383
	3	0	0,404
8013	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-1793,43	1
	2	0,4497198	0,295
	3	0,4497198	0,295
8023	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,6339341	0,268
	2	-1978,66	1
	3	1,0897103	0,175
8032	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,9073473	0,2
	2	1,3123747	0,1
	3	-4775,66	1
8073	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,0715308	0,481
	2	-3536,16	1
	3	0,0715308	0,481
8075	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,5829666	0,265
	2	-2078,01	1
	3	0,6690581	0,269

<i>maxgen</i>			
7316	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-865,36	1
	2	0,5546284	0,282
	3	0,5546829	0,282
7339	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-411,01	1
	2	0	0,143
	3	0	0,149
7383	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0	0,577
	2	0	0,256
	3	-422,69	1
7434	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,6926182	0,233
	2	-1622,66	1
	3	0,6383818	0,239
7512	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-3521,72	1
	2	1,019828	0,146
	3	1,0324061	0,144
7538	Tree	$\Delta\log L$ :S.E.	p-SH
	1	3,1920693	<0.000
	2	3,1920288	<0.000
	3	-4515,29	1
7606	Tree	$\Delta\log L$ :S.E.	p-SH
	1	1,1938996	0,156
	2	0,5400576	0,317
	3	-2060,65	1
7631	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-3929,51	1
	2	0,8806862	0,182
	3	0,7720076	0,202
7640	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0	0,469
	2	-3387,26	1
	3	0	0,002
7720	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,7095904	0,242
	2	0,7095904	0,242
	3	-1405,59	1
7736	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,0459175	0,588
	2	-1684,04	1
	3	0,6702743	0,432
7742	Tree	$\Delta\log L$ :S.E.	p-SH
	1	1,2103654	0,111
	2	-1732,85	1
	3	1,2104043	0,111
7771	Tree	$\Delta\log L$ :S.E.	p-SH
	1	1,2290534	0,108
	2	-1083,55	1
	3	1,2290534	0,108
7864	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,2301055	0,415
	2	-1576,5	1
	3	0,2215657	0,4

<i>maxspe</i>			
<b>8076</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0	0,576
	2	0	0,688
	3	-2968,41	1
<b>8456</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0	1
	2	0	0,505
	3	-1180,76	0,321
<b>8547</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,209205	0,382
	2	-1012,64	1
	3	0,2092676	0,383
<b>8671</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,5369873	0,298
	2	0,5368879	0,298
	3	-6150,07	1
<b>8732</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0	0,599
	2	0	0,467
	3	-2790,14	1
<b>8784</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0	0,022
	2	-592,62	1
	3	0	0,02
<b>8997</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,0895996	0,573
	2	-1917,3	1
	3	0,4004843	0,532
<b>9404</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,117325	0,435
	2	-817,29	1
	3	0,1172447	0,435
<b>9414</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,3289474	0,339
	2	0,3289474	0,339
	3	-708,87	1
<b>9562</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	1,150904	0,116
	2	-7451,58	1
	3	1,1867649	0,12
<b>9590</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	1,0229481	0,143
	2	-3019,65	1
	3	0,7899571	0,203
<b>9821</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	1,0445682	0,153
	2	-1579,59	1
	3	1,0118905	0,166
<b>9827</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	-1136,47	1
	2	0,7515325	0,228
	3	0,7515009	0,228

<i>maxgen</i>			
<b>7883</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,2663559	0,375
	2	0,2662673	0,375
	3	-1437,09	1
<b>7902</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,0388686	0,609
	2	1,2289928	0,231
	3	-2242,52	1
<b>7950</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,4429436	0,475
	2	0,3309341	0,499
	3	-816,95	1
<b>7970</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	-444,18	1
	2	0,5630301	0,27
	3	0,5630713	0,27
<b>8013</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	-1285,73	1
	2	0,7215007	0,207
	3	0,7204611	0,205
<b>8023</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,0039897	0,63
	2	-1306,04	1
	3	1,0660395	0,263
<b>8032</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,6541799	0,253
	2	0,6048119	0,261
	3	-3139,07	1
<b>8073</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0	0,775
	2	-2410,92	1
	3	0	0,571
<b>8075</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,3565062	0,333
	2	-1319,4	1
	3	0,6096025	0,27
<b>8076</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0	0,595
	2	-2301,82	1
	3	0	0,669
<b>8092</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,505771	0,291
	2	-2621,95	1
	3	0,606147	0,284
<b>8323</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	-3753,77	1
	2	1,6241299	0,08
	3	1,6298021	0,079
<b>8671</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	-4482,29	1
	2	0,729064	0,228
	3	0,7289683	0,228
<b>8732</b>	Tree	$\Delta\log L:S.E.$	p-SH
	1	0,096432	0,452
	2	0,0964785	0,451
	3	-1933,57	1

*maxgen*

<b>8782</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,5064045	0,274
	2	0,5064422	0,274
	3	-830,38	1
<b>8784</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0	0,123
	2	-528,5	1
	3	0	0,088
<b>8942</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,7085586	0,221
	2	-941,32	1
	3	0,7085917	0,221
<b>8997</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-1278,6	1
	2	0,4153611	0,366
	3	0,566242	0,35
<b>9007</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	1,3919881	0,091
	2	-1304,15	1
	3	1,3053976	0,103
<b>9095</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,8640949	0,204
	2	1,5030554	0,081
	3	-2197,76	1
<b>9169</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,3037712	0,501
	2	-1596,79	1
	3	0,266609	0,528
<b>9384</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	1,3762396	0,085
	2	1,3425845	0,089
	3	-855,32	1
<b>9414</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0	0,153
	2	0	0,181
	3	-544,6	1
<b>9489</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	7,3617103	<0.000
	2	-2754,7	1
	3	7,3617103	<0.000
<b>9511</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	11,596094	<0.000
	2	-3583,33	1
	3	11,596094	<0.000
<b>9562</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,7235232	0,228
	2	-4877,63	1
	3	0,8860663	0,19
<b>9569</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,7427612	0,235
	2	-1300,01	1
	3	0,7882245	0,219
<b>9590</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	1,2501223	0,107
	2	-1982,62	1
	3	1,1124446	0,135

*maxgen*

<b>9594</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	1,3597108	0,099
	2	1,5247693	0,075
	3	-4281,75	1
<b>9616</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	0,36	0,36
	2	-2560,89	1
	3	0,359952	0,36
<b>9813</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-765,87	1
	2	0,830391	0,179
	3	0,7897706	0,193
<b>9827</b>	Tree	$\Delta\log L$ :S.E.	p-SH
	1	-846,49	1
	2	0,5775468	0,244
	3	0,5775211	0,244

**Supplementary table S4** – Maximum likelihood bootstrap support of individual alignments (assigned with the numerical identifier).

For each gene alignment of the maxspe set (Table S4a) and of the maxgen set (Table S4b) a maximum likelihood tree with 100 bootstrap replicates was calculated using RAXML. The first column refers to the gene ID of HaMStR, the second column indicates the tree topology. If the topology coincides with a concurrent hypothesis (Palaeoptera, Metapterygota, Chiasmomyaria) the bootstrap value of the branch that separates the respective out-group from the respective in-group is displayed.

**Table S4a maxspe**

ID Number	Topology	Relevant Bootstrap
6671	Other Topology	
6790	Other Topology	
6906	Other Topology	
6927	Other Topology	
6958	Other Topology	
7007	Other Topology	
7015	Other Topology	
7098	Other Topology	
7214	Other Topology	
7316	Other Topology	
7339	Other Topology	
7434	Other Topology	
7512	Other Topology	
7538	Other Topology	
7606	Other Topology	
7631	Other Topology	
7640	Other Topology	
7720	Other Topology	
7883	Other Topology	
7950	Chiasmomyaria	46
7970	Other Topology	
8013	Other Topology	
8023	Metapterygota	34
8032	Other Topology	
8073	Other Topology	
8075	Metapterygota	52
8076	Other Topology	
8456	Other Topology	
8547	Other Topology	
8671	Other Topology	
8732	Palaeoptera	18
8784	Other Topology	
8997	Other Topology	
9404	Other Topology	
9414	Other Topology	
9562	Other Topology	
9590	Other Topology	
9821	Other Topology	
9827	Other Topology	

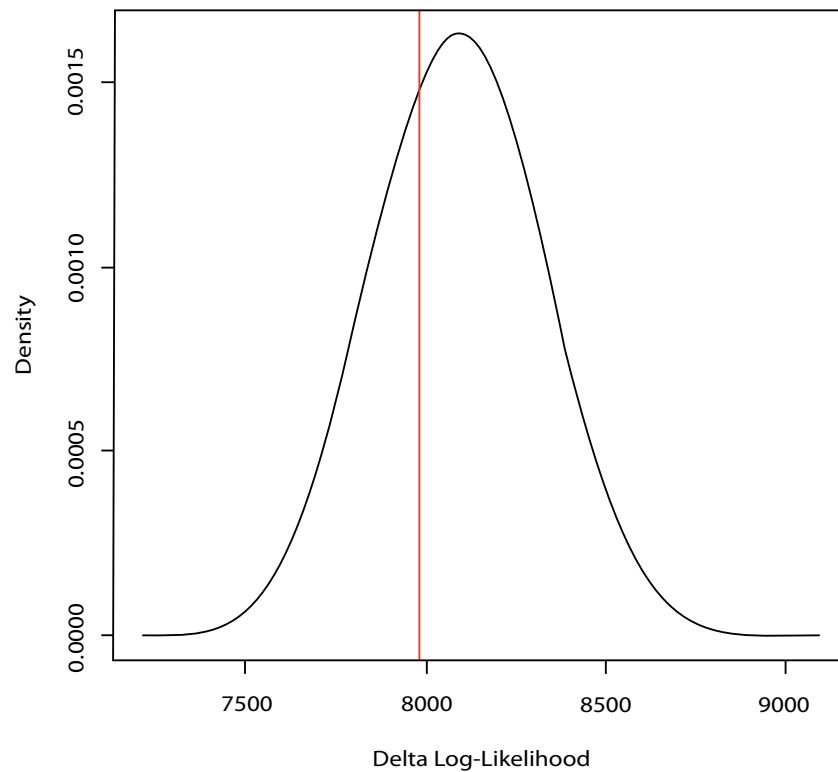
**Table S4b maxgen**

GeneID	Topology	Relevant Bootstrap
6637	Other Topology	
6671	Other Topology	
6715	Other Topology	
6790	Other Topology	
6906	Other Topology	
6927	Other Topology	
6936	Other Topology	
6958	Other Topology	
7007	Other Topology	
7015	Palaeoptera	45
7098	Other Topology	
7214	Other Topology	
7316	Other Topology	
7339	Other Topology	
7383	Palaeoptera	5
7434	Other Topology	
7512	Other Topology	
7538	Other Topology	
7606	Other Topology	
7631	Other Topology	
7640	Other Topology	
7720	Other Topology	
7736	Other Topology	
7742	Metapterygota	50
7771	Other Topology	
7864	Other Topology	
7883	Chiasmomyaria	35
7902	Palaeoptera	24
7950	Chiasmomyaria	56
7970	Other Topology	
8013	Other Topology	
8023	Other Topology	
8032	Other Topology	
8073	Other Topology	
8075	Metapterygota	63
8076	Other Topology	
8092	Palaeoptera	61
8323	Other Topology	
8671	Palaeoptera	59
8732	Other Topology	
8782	Other Topology	
8784	Other Topology	
8942	Metapterygota	41
8997	Other Topology	
9007	Other Topology	
9095	Other Topology	
9169	Other Topology	
9384	Other Topology	
9414	Other Topology	
9489	Other Topology	
9511	Other Topology	
9562	Other Topology	
9569	Other Topology	
9590	Metapterygota	95
9594	Other Topology	
9616	Other Topology	
9813	Other Topology	
9827	Other Topology	

**Supplementary Figure S1** – Distribution of delta values for simulated *maxgen* alignments without gaps.

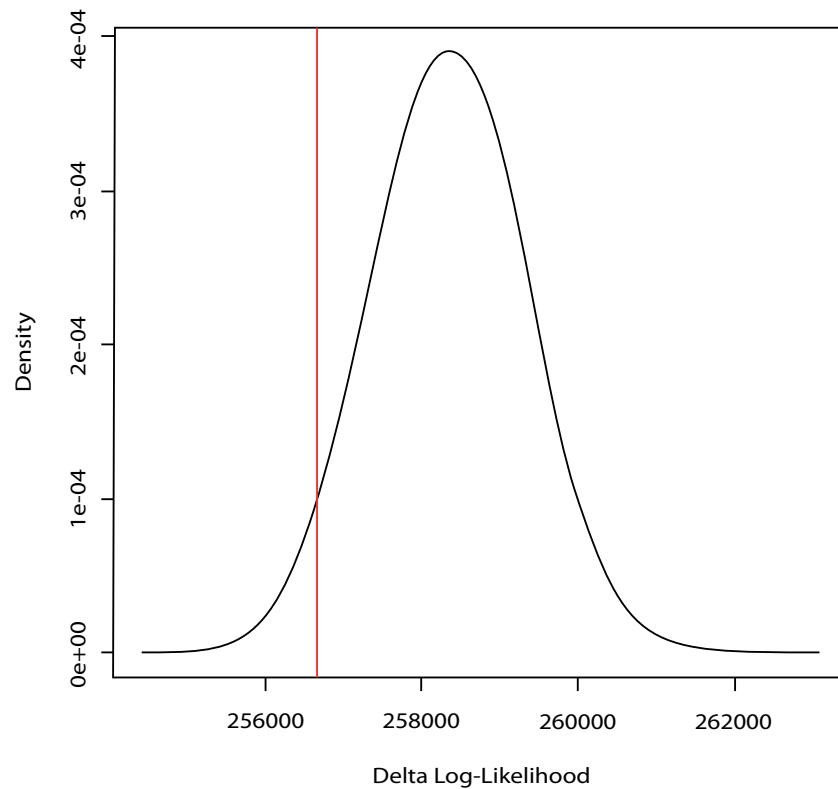
We removed all columns containing gaps from the concatenated *maxgen* alignment processed with ALISCORE and calculated a maximum likelihood tree with RAxML under the WAG model. Afterwards we simulated 1,000 alignments of equal length using Seq-Gen and the parameters obtained by the maximum likelihood tree reconstruction. Following the test introduced in Goldman (1993) we reconstructed the maximum likelihood tree and calculated the difference of the unconstrained log-likelihood and the maximum log-likelihood (delta value) for each simulated alignment.

Shown is the distribution of delta values for the simulated alignments. The red vertical line marks the delta value for the real alignment.



**Supplementary Figure S2** – Distribution of delta values for simulated *maxgen* alignments with gaps.

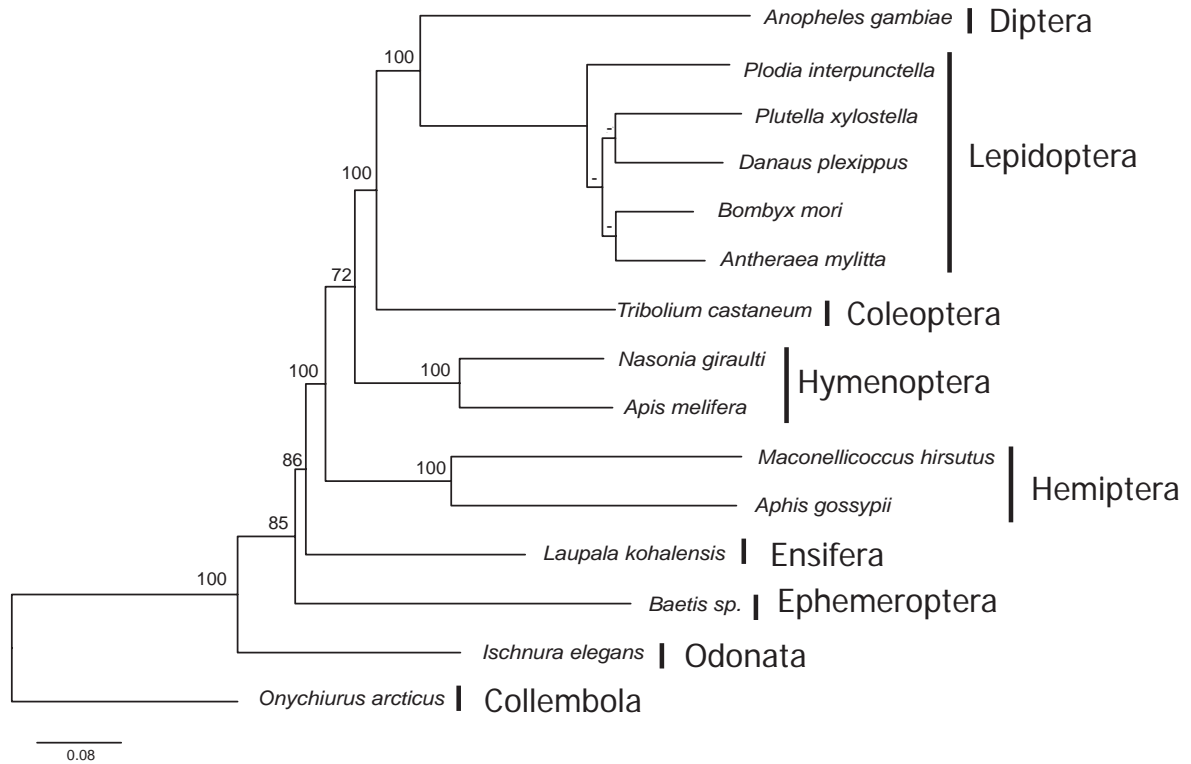
We simulated 1,000 alignments with Seq-Gen using the parameter obtained by the maximum likelihood tree reconstruction of the concatenated *maxgen* alignment including all positions with gaps or missing data. We then replaced amino acids of the simulated data with gaps or missing data where there are gaps or missing data in the real alignment. Afterwards we proceeded as described in Supplementary Figure S1. Shown is the distribution of delta values of the simulated alignments. The red vertical line marks the delta value for the real alignment.



**Supplementary Figure S3** – Maximum likelihood topology of *maxspe*

Each gene alignment of the *maxspe* set was processed with ALISCORE. Afterwards the best suited model of evolution was determined for each processed alignment with ProtTest and the alignments were concatenated. The maximum likelihood tree was calculated using RAxML's '-q' option, that allows a partitioning of the alignment with an individual model of evolution for each partition. Support values were assessed by 100 bootstrap replicates.

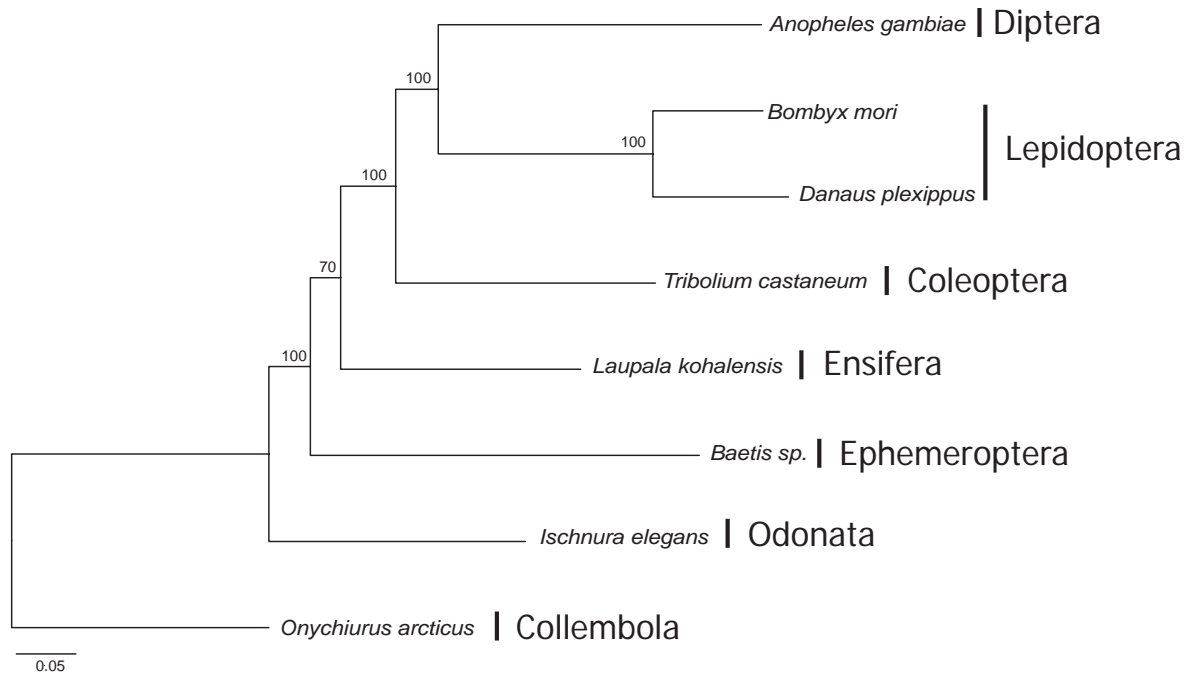
The tree is congruent to the tree shown in Figure 2. Thus, even if sequence evolution is modeled individually for each gene, the Chiasmomyaria hypothesis is supported.





**Supplementary Figure S4** – Maximum likelihood topology of *maxgen*

Each gene alignment of the *maxgen* set was processed with ALISCORE. Afterwards the best suited model of evolution was determined for each with ProtTest and the alignments were concatenated. The maximum likelihood tree was calculated using RAxML's '-q' option, that allows a partitioning of the alignment with an individual model of evolution for each partition. Support values were assessed by 100 bootstrap replicates. The tree is congruent to the tree shown in Figure 3. Thus, even if sequence evolution is modeled individually for each gene, the Chiasmomyaria hypothesis is supported.



### Supplementary Table 1 | Taxa included in analyses.

Species written in capitals – used proteome species; # Taxa included in the optimal data subset selected by reduction heuristics; #Taxa used to train Hidden Markov Models (HMMs) to predict putative orthologs<sup>1</sup>; Source: dbEST – <http://www.ncbi.nlm.nih.gov/dbEST>, Gene Index Project (gendix) – <http://compbi.dfci.harvard.edu/tgi/tgipage.html>, NCBI Trace Archive – <ftp://ftp.ncbi.nih.gov/pub/TraceDB>, JGI –<http://www.jgi.doe.gov>, InParanoid<sup>2,3</sup> v6.1<sup>4</sup> – <http://inparanoid6.sbc.su.se>, VectorBase –<http://www.vectorbase.org>, BeetleBase – <http://beetlebase.org/>, SilkDB – <http://silkworm.genomics.org.cn/>, UniProt (integr8) – <http://www.ebi.ac.uk/integr8/>, UCSC – <http://hgdownload.cse.ucsc.edu>; Data of *C. pipiens quinquefasciatus* was kindly provided by the Broad Institute of MIT and Harvard (USA); No. of EST contigs – number of assembled EST contigs; No. of genes orig. data set – number of orthologous genes per taxon in the original data set; No. of genes data subset – number of orthologous genes per taxon in the optimal data subset after performing reduction heuristics.

Species	Group	Source	No. of EST contigs	No. of genes orig. data set	No. of genes data subset
<i>Hypsibius dujardini</i> <sup>#</sup>	Panarthropoda, Tardigrada	dbEST	2,386	140	81
<i>Richtersius coronifer</i> <sup>#</sup>	Panarthropoda, Tardigrada	NCBI Trace Archive	1,537	99	52
<i>Epiperipatus</i> sp. TB-2001	Panarthropoda, Onychophora	dbEST	825	49	
<i>Peripatopsis sedgwicki</i> <sup>#</sup>	Panarthropoda, Onychophora	present study	3,452	142	72
<i>Euperipatoides kanangrensis</i> <sup>#</sup>	Panarthropoda, Onychophora	NCBI Trace Archive	1,449	110	53
<i>Julida</i> sp. APV-2005	Arthropoda, Myriapoda	dbEST	231	13	
<i>Archispirostreptus gigas</i> <sup>#</sup>	Arthropoda, Myriapoda	present study	2,299	117	58
<i>Scutigera coleoptrata</i> <sup>#</sup>	Arthropoda, Myriapoda	NCBI Trace Archive	807	54	35
<i>Anoplodactylus eroticus</i> <sup>#</sup>	Arthropoda, Chelicerata	NCBI Trace Archive	1,281	91	55
<i>Endeis spinosa</i> <sup>#</sup>	Arthropoda, Chelicerata	present study	2,672	174	69
<i>Limulus polyphemus</i> <sup>#</sup>	Arthropoda, Chelicerata	present study	4,050	210	89
<i>Carcinoscorpius rotundicauda</i>	Arthropoda, Chelicerata	dbEST	512	21	
<i>Mesobuthus gibbosus</i>	Arthropoda, Chelicerata	dbEST	587	38	
<i>Loxosceles laeta</i>	Arthropoda, Chelicerata	dbEST	1,209	66	
<i>Dysdera erythrina</i>	Arthropoda, Chelicerata	dbEST	279	22	
<i>Cupiennius salei</i>	Arthropoda, Chelicerata	dbEST	208	30	
<i>Araneus ventricosus</i>	Arthropoda, Chelicerata	dbEST	204	11	
<i>Acanthoscurria gomesiana</i> <sup>#</sup>	Arthropoda, Chelicerata	dbEST	3,713	234	90
<i>Chilobrachys jingzhao</i>	Arthropoda, Chelicerata	dbEST	230	22	
<i>Ixodes scapularis</i> <sup>#</sup>	Arthropoda, Chelicerata	genidx	38,275	578	128
<i>Ixodes ricinus</i>	Arthropoda, Chelicerata	dbEST	1,300	53	
<i>Amblyomma variegatum</i> <sup>#</sup>	Arthropoda, Chelicerata	genidx	2,109	162	62
<i>Amblyomma americanum</i> <sup>#</sup>	Arthropoda, Chelicerata	dbEST	2,798	88	44
<i>Amblyomma cajennense</i>	Arthropoda, Chelicerata	dbEST	1,165	71	
<i>Dermacentor andersoni</i> <sup>#</sup>	Arthropoda, Chelicerata	dbEST	752	63	38
<i>Dermacentor variabilis</i>	Arthropoda, Chelicerata	dbEST	1,075	49	
<i>Boophilus microplus</i> <sup>#</sup>	Arthropoda, Chelicerata	dbEST	14,507	425	112
<i>Rhipicephalus appendiculatus</i> <sup>#</sup>	Arthropoda, Chelicerata	genidx	7,359	321	92
<i>Argas monolakensis</i>	Arthropoda, Chelicerata	dbEST	1,620	51	
<i>Ornithodoros porcinus porcinus</i>	Arthropoda, Chelicerata	dbEST	771	29	
<i>Ornithodoros parkeri</i>	Arthropoda, Chelicerata	dbEST	689	37	
<i>Ornithodoros coriaceus</i>	Arthropoda, Chelicerata	dbEST	702	19	
<i>Glycyphagus domesticus</i> <sup>#</sup>	Arthropoda, Chelicerata	dbEST	2,511	97	56
<i>Blomia tropicalis</i> <sup>#</sup>	Arthropoda, Chelicerata	dbEST	1,331	80	37
<i>Psoroptes ovis</i>	Arthropoda, Chelicerata	dbEST	281	18	
<i>Sarcoptes scabiei</i>	Arthropoda, Chelicerata	dbEST	817	38	
<i>Dermatophagoides pteronyssinus</i>	Arthropoda, Chelicerata	dbEST	1,258	67	
<i>Dermatophagoides farinae</i>	Arthropoda, Chelicerata	dbEST	1,046	59	
<i>Suidasia medanensis</i> <sup>#</sup>	Arthropoda, Chelicerata	dbEST	2,083	139	73
<i>Tyrophagus putrescentiae</i>	Arthropoda, Chelicerata	dbEST	881	46	
<i>Acarus siro</i>	Arthropoda, Chelicerata	dbEST	652	57	
<i>Aleuroglyphus ovatus</i>	Arthropoda, Chelicerata	dbEST	1,440	58	
<i>Gammarus pulex</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	4,241	102	63
<i>Eurydice pulchra</i>	Arthropoda, Crustacea	dbEST	562	26	
<i>Euphausia superba</i>	Arthropoda, Crustacea	dbEST	1,101	43	
<i>Homarus americanus</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	14,147	383	111

# Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

<i>Pacifastacus leniusculus</i>	Arthropoda, Crustacea	dbEST	175	14	
<i>Petrolisthes cinctipes</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	27,086	416	119
<i>Callinectes sapidus</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	2,239	114	56
<i>Carcinus maenas</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	4,567	233	76
<i>Cancer magister</i>	Arthropoda, Crustacea	dbEST	445	14	
<i>Celca pugilator</i>	Arthropoda, Crustacea	dbEST	1,482	64	
<i>Gecarcoidea natalis</i>	Arthropoda, Crustacea	dbEST	656	23	
<i>Ilyoplax pusilla</i>	Arthropoda, Crustacea	dbEST	251	2	
<i>Eriocheir sinensis</i>	Arthropoda, Crustacea	dbEST	1,136	58	
<i>Marsupenaeus japonicus</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	1,944	61	46
<i>Fenneropenaeus chinensis</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	3,458	114	74
<i>Penaeus monodon</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	4,097	129	81
<i>Litopenaeus vannamei</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	3,774	126	75
<i>Litopenaeus stylirostris</i>	Arthropoda, Crustacea	dbEST	314	12	
<i>Litopenaeus setiferus</i>	Arthropoda, Crustacea	dbEST	642	50	
<i>Tigriopus californicus</i> <sup>#</sup>	Arthropoda, Crustacea	present study	2,598	65	39
<i>Calanus finmarchicus</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	4,906	189	49
<i>Lepeophtheirus salmonis</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	5,102	339	98
<i>Pollicipes pollicipes</i> <sup>#</sup>	Arthropoda, Crustacea	present study	1,721	107	59
<i>Artemia franciscana</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	10,330	323	116
<i>Triops cancriformis</i> <sup>#</sup>	Arthropoda, Crustacea	present study	2,542	115	54
<i>Daphnia magna</i> <sup>#</sup>	Arthropoda, Crustacea	dbEST	5,307	207	85
DAPHNIA PULEX <sup>#, §</sup>	Arthropoda, Crustacea	JGI	30,939	775	129
<i>Acerentomon franzi</i> <sup>#</sup>	Arthropoda, Hexapoda	present study	1,995	99	52
<i>Campodea cf. fragilis</i> <sup>#</sup>	Arthropoda, Hexapoda	present study	6,407	150	68
<i>Folsomia candida</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	5,955	143	41
<i>Anurida maritima</i> <sup>#</sup>	Arthropoda, Hexapoda	present study	3,504	131	53
<i>Onychiurus arcticus</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	9,981	309	89
<i>Lepismachilis y-signata</i> <sup>#</sup>	Arthropoda, Hexapoda	present study	2,288	123	66
<i>Tricholepisma aurea</i>	Arthropoda, Hexapoda	dbEST	344	34	
<i>Ischnura elegans</i> <sup>#</sup>	Arthropoda, Hexapoda	Simon et al. (2009)	3,194	177	66
<i>Baetis</i> sp. <sup>#</sup>	Arthropoda, Hexapoda	Simon et al. (2009)	3,035	144	49
<i>Locusta migratoria</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	12,255	303	107
<i>Allonemobius fasciatus</i>	Arthropoda, Hexapoda	dbEST	116	10	
<i>Laupala kohalensis</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	8,371	292	90
<i>Gryllus bimaculatus</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	3,945	238	93
<i>Gryllus pennsylvanicus</i>	Arthropoda, Hexapoda	dbEST	338	30	
<i>Gryllus firmus</i>	Arthropoda, Hexapoda	dbEST	271	14	
<i>Periplaneta americana</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,577	84	58
<i>Blattella germanica</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,546	75	38
<i>Diploptera punctata</i>	Arthropoda, Hexapoda	dbEST	666	20	
<i>Hodotermopsis sjostedti</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,471	73	46
<i>Reticulitermes flavipes</i>	Arthropoda, Hexapoda	dbEST	113	1	
<i>Sphodromantis centralis</i>	Arthropoda, Hexapoda	dbEST	120	4	
PEDICULUS HUMANUS <sup>#</sup>	Arthropoda, Hexapoda	VectorBase	11,198	636	122
<i>Pediculus humanus corporis</i>	Arthropoda, Hexapoda	dbEST	472	55	
<i>Pediculus humanus capitis</i>	Arthropoda, Hexapoda	dbEST	2,868	147	
<i>Homalodisca coagulata</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	5,661	237	96
<i>Graphocephala atropunctata</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,827	97	63
<i>Oncometopia nigricans</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,772	114	63
<i>Lygus lineolaris</i>	Arthropoda, Hexapoda	dbEST	371	21	
<i>Oncopeltus fasciatus</i>	Arthropoda, Hexapoda	dbEST	448	11	
<i>Rhodnius prolixus</i>	Arthropoda, Hexapoda	dbEST	735	48	
<i>Triatoma infestans</i>	Arthropoda, Hexapoda	dbEST	908	39	
<i>Triatoma brasiliensis</i>	Arthropoda, Hexapoda	dbEST	1,897	33	
<i>Bemisia tabaci</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	4,548	61	40
<i>Aleurothrixus</i> sp. APV-2005	Arthropoda, Hexapoda	dbEST	288	18	
<i>Pachypsylla venusta</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	4,631	118	56
<i>Diaphorina citri</i>	Arthropoda, Hexapoda	dbEST	2,257	66	
<i>Aphis gossypii</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	3,716	210	88
<i>Myzus persicae</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	9,946	447	107
<i>Acyrtosiphon pisum</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	18,253	413	110
<i>Rhopalosiphum padi</i>	Arthropoda, Hexapoda	dbEST	335	34	
<i>Toxoptera citricida</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	2,196	143	74

## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

<i>Sogatella furcifera</i>	Arthropoda, Hexapoda	dbEST	122	9	
<i>Nilaparvata lugens</i>	Arthropoda, Hexapoda	dbEST	167	7	
<i>Maconellicoccus hirsutus</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	3,929	217	85
<i>Nasonia giraulti</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	6,764	277	101
<i>Nasonia vitripennis</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	2,999	160	86
<i>Copidosoma floridanum</i>	Arthropoda, Hexapoda	dbEST	216	9	
<i>Lysiphlebus testaceipes</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	3,881	210	84
<i>Microctonus hyperodae</i>	Arthropoda, Hexapoda	dbEST	545	22	
<i>Vespula squamosa</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,227	70	50
<i>Solenopsis invicta</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	12,252	297	95
<i>Camponotus festinatus</i>	Arthropoda, Hexapoda	dbEST	149	8	
<i>Lasius niger</i>	Arthropoda, Hexapoda	dbEST	347	3	
<i>Bombus ignitus</i>	Arthropoda, Hexapoda	dbEST	213	22	
<i>APIS MELLIFERA</i> <sup>#,§</sup>	Arthropoda, Hexapoda	InParanoid	13,448	775	129
<i>Melipona quadrifasciata</i>	Arthropoda, Hexapoda	dbEST	321	2	
<i>Eoxenos laboulbenei</i>	Arthropoda, Hexapoda	dbEST	345	32	
<i>Mengenilla chobauti</i>	Arthropoda, Hexapoda	dbEST	297	27	
<i>Micromalthus debilis</i>	Arthropoda, Hexapoda	dbEST	157	13	
<i>Carabus granulatus</i>	Arthropoda, Hexapoda	dbEST	177	16	
<i>Meladema coriacea</i>	Arthropoda, Hexapoda	dbEST	328	23	
<i>Cicindela littorea</i>	Arthropoda, Hexapoda	dbEST	232	5	
<i>Cicindela campestris</i>	Arthropoda, Hexapoda	dbEST	340	24	
<i>Cicindela littoralis</i>	Arthropoda, Hexapoda	dbEST	236	12	
<i>Sphaerius</i> sp. APV-2005	Arthropoda, Hexapoda	dbEST	396	29	
<i>Eucinetus</i> sp. APV-2005	Arthropoda, Hexapoda	dbEST	344	27	
<i>Dascillus cervinus</i>	Arthropoda, Hexapoda	dbEST	354	28	
<i>Georissus</i> sp. APV-2005	Arthropoda, Hexapoda	dbEST	408	33	
<i>Trox</i> sp. JH-2005	Arthropoda, Hexapoda	dbEST	223	9	
<i>Scarabaeus laticollis</i>	Arthropoda, Hexapoda	dbEST	328	30	
<i>Julodis onopordi</i>	Arthropoda, Hexapoda	dbEST	337	24	
<i>Hister</i> sp. APV-2005	Arthropoda, Hexapoda	dbEST	358	35	
<i>Agriotes lineatus</i>	Arthropoda, Hexapoda	dbEST	452	22	
<i>Tenebrio molitor</i>	Arthropoda, Hexapoda	dbEST	100	3	
<i>TRIBOLIUM CASTANEUM</i> <sup>#,§</sup>	Arthropoda, Hexapoda	BeetleBase	16,421	775	129
<i>Mycetophagus quadripustulatus</i>	Arthropoda, Hexapoda	dbEST	419	28	
<i>Biphyllus lunatus</i>	Arthropoda, Hexapoda	dbEST	260	28	
<i>Hypothenemus hampei</i>	Arthropoda, Hexapoda	dbEST	844	64	
<i>Diaprepes abbreviatus</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,921	65	42
<i>Curculio glandium</i>	Arthropoda, Hexapoda	dbEST	241	25	
<i>Sitophilus zeamais</i>	Arthropoda, Hexapoda	dbEST	82	8	
<i>Ips pini</i>	Arthropoda, Hexapoda	dbEST	565	58	
<i>Platystomus albinus</i>	Arthropoda, Hexapoda	dbEST	145	5	
<i>Diabrotica virgifera virgifera</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	7,871	336	114
<i>Timarcha balearica</i>	Arthropoda, Hexapoda	dbEST	272	21	
<i>Leptinotarsa decemlineata</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	2,668	122	56
<i>Callosobruchus maculatus</i>	Arthropoda, Hexapoda	dbEST	561	58	
<i>Anoplophora glabripennis</i>	Arthropoda, Hexapoda	dbEST	386	31	
<i>Limnephilus flavicornis</i>	Arthropoda, Hexapoda	dbEST	117	2	
<i>Hydropsyche</i> sp. T20	Arthropoda, Hexapoda	dbEST	203	23	
<i>Plutella xylostella</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,048	72	55
<i>Tineola bisselliella</i>	Arthropoda, Hexapoda	dbEST	188	7	
<i>Danaus plexippus</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	9,930	470	114
<i>Bicyclus anynana</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	5,575	165	68
<i>Heliconius erato</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	3,327	219	93
<i>Heliconius melpomene</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,820	104	64
<i>Papilio dardanus</i>	Arthropoda, Hexapoda	dbEST	310	52	
<i>Plodia interpunctella</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	3,808	175	81
<i>Ostrinia nubilalis</i>	Arthropoda, Hexapoda	dbEST	489	25	
<i>Epiphyas postvittana</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	2,895	154	88
<i>Choristoneura fumiferana</i>	Arthropoda, Hexapoda	dbEST	589	17	
<i>Trichoplusia ni</i>	Arthropoda, Hexapoda	dbEST	417	42	
<i>Agrotis segetum</i>	Arthropoda, Hexapoda	dbEST	812	58	
<i>Spodoptera litura</i>	Arthropoda, Hexapoda	dbEST	61	3	
<i>Spodoptera frugiperda</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	8,362	309	123

# Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

<i>Heliothis virescens</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,723	167	73
<i>Helicoverpa armigera</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	692	70	54
<i>Euclidia glyphica</i>	Arthropoda, Hexapoda	dbEST	187	16	
<i>Bombyx mandarina</i>	Arthropoda, Hexapoda	dbEST	207	12	
<i>BOMBYX MORI</i> <sup>#, §</sup>	Arthropoda, Hexapoda	SilkDB	16,329	775	129
<i>Manduca sexta</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	2,197	120	68
<i>Lonomia obliqua</i>	Arthropoda, Hexapoda	dbEST	610	58	
<i>Samia cynthia ricini</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	5,721	254	105
<i>Antheraea yamamai</i>	Arthropoda, Hexapoda	dbEST	421	27	
<i>Antheraea assama</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	8,927	292	108
<i>Antheraea mylitta</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,478	93	58
<i>Panorpa cf. vulgaris</i> APV-2005	Arthropoda, Hexapoda	dbEST	322	21	
<i>Ctenocephalides felis</i>	Arthropoda, Hexapoda	dbEST	1,775	82	
<i>Xenopsylla cheopis</i>	Arthropoda, Hexapoda	dbEST	283	26	
<i>Culicoides sonorensis</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,405	90	62
<i>Chironomus tentans</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	3,445	216	97
<i>ANOPHELES GAMBIAE</i> <sup>#</sup>	Arthropoda, Hexapoda	UniProt (integr8)	12,463	726	126
<i>Anopheles aquasalis</i>	Arthropoda, Hexapoda	dbEST	121	4	
<i>Anopheles darlingi</i>	Arthropoda, Hexapoda	dbEST	461	24	
<i>Anopheles albimanus</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	3,096	94	53
<i>Anopheles anthropophagus</i>	Arthropoda, Hexapoda	dbEST	141	5	
<i>Anopheles funestus</i>	Arthropoda, Hexapoda	dbEST	1,224	59	
<i>AEDES AEGYPTI</i> <sup>#, §</sup>	Arthropoda, Hexapoda	InParanoid	15,419	654	112
<i>Armigeres subalbatus</i> <sup>#</sup>	Arthropoda, Hexapoda	NCBI Trace Archive	7,770	329	97
<i>CULEX PIPPIENS QUINQUEFASCIATUS</i> <sup>#</sup>	Arthropoda, Hexapoda	Broad Institute	20,306	721	128
<i>Culex pipiens pallens</i>	Arthropoda, Hexapoda	dbEST	76	3	
<i>Toxorhynchites amboinensis</i>	Arthropoda, Hexapoda	dbEST	199	7	
<i>Lutzomyia longipalpis</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	19,739	478	126
<i>Phlebotomus papatasi</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	10,797	422	125
<i>Rhynchosciara americana</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	3,449	112	66
<i>Mayetiola destructor</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	1,482	81	48
<i>Sitodiplosis mosellana</i>	Arthropoda, Hexapoda	dbEST	1,100	64	
<i>Orseolia oryzae</i>	Arthropoda, Hexapoda	dbEST	976	29	
<i>Glossina morsitans morsitans</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	12,444	512	124
<i>Musca domestica</i>	Arthropoda, Hexapoda	dbEST	296	14	
<i>Stomoxys calcitrans</i>	Arthropoda, Hexapoda	dbEST	296	31	
<i>Haematobia irritans</i>	Arthropoda, Hexapoda	dbEST	196	13	
<i>Haematobia irritans irritans</i>	Arthropoda, Hexapoda	dbEST	189	13	
<i>Ceratitis capitata</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	11,132	475	123
<i>Rhagoletis suavis</i>	Arthropoda, Hexapoda	dbEST	370	27	
<i>Rhagoletis pomonella</i>	Arthropoda, Hexapoda	dbEST	160	7	
<i>Drosophila arizonae</i> <sup>#</sup>	Arthropoda, Hexapoda	dbEST	770	88	55
<i>DROSOPHILA ANANASSAE</i> <sup>#</sup>	Arthropoda, Hexapoda	USCS	29,704	673	113
<i>DROSOPHILA ERECTA</i> <sup>#</sup>	Arthropoda, Hexapoda	USCS	17,531	673	117
<i>DROSOPHILA MELANOGASTER</i> <sup>#, §</sup>	Arthropoda, Hexapoda	InParanoid	13,854	752	129
<i>Meloidogyne hapla</i> <sup>#</sup>	Nematoda	dbEST	7,802	252	92
<i>CAENORHABDITIS ELEGANS</i> <sup>#, §</sup>	Nematoda	InParanoid	20,084	749	127
<i>CAENORHABDITIS REMANEI</i> <sup>#</sup>	Nematoda	InParanoid	25,595	719	126
<i>CAENORHABDITIS BRIGGSIAE</i> <sup>#, §</sup>	Nematoda	InParanoid	19,334	711	126
<i>Haemonchus contortus</i> <sup>#</sup>	Nematoda	dbEST	5,842	262	98
<i>Ascaris suum</i> <sup>#</sup>	Nematoda	dbEST	9,165	197	84
<i>Xiphinema index</i> <sup>#</sup>	Nematoda	dbEST	4,824	228	89
<i>Trichinella spiralis</i> <sup>#</sup>	Nematoda	dbEST	8,843	373	111
<i>CAPITELLA CAPITATA</i> <sup>#, §</sup>	Annelida	JGI	32,415	724	122
<i>HELOBDELLA ROBUSTA</i> <sup>#</sup>	Annelida	JGI	23,432	730	126
<i>Lumbricus rubellus</i> <sup>#</sup>	Annelida	dbEST	10,386	196	94
<i>LOTTIA GIGANTEA</i> <sup>#, §</sup>	Mollusca	JGI	23,851	672	120
<i>Crassostrea gigas</i> <sup>#</sup>	Mollusca	dbEST	14,857	339	102
<i>Argopecten irradians</i> <sup>#</sup>	Mollusca	dbEST	3,610	95	59

# Supplementary Table 2 | Genes selected by HaMStR<sup>1</sup> and used in phylogenetic analyses.

Gene ID– numerical internal identifier that corresponds to the partition number (gene number) of the data matrix; Protein ID – FlyBase-ID from Ensembl Archive February 2007 (Ensembl Arch. 02/07) for *Drosophila melanogaster*, <http://feb2007.archive.ensembl.org/> respectively AEE-ID from Inparanoid<sup>4</sup> v6.1 for *Aede saegypti*, <http://inparanoid6.sbc.su.se>; Gene / Description – Description of genes as determined from the Ensembl Archive / Flybase for *D. melanogaster* (Dmel), from InParanoid v6.1 for *A. aegypti* (Aaeg) or from *HomoloGene* for *Homo sapiens* (Hsap), <http://www.ncbi.nlm.nih.gov/homologene>; other studies –genes shared with other studies: ph – Philippe *et al.*<sup>5</sup>; de – Delsuc *et al.*<sup>6</sup>; du – Dunn *et al.*<sup>7</sup>; ba – Baurain *et al.*<sup>8</sup>; name assigned to the genes in previous studies are given in squared brackets; Rib. Protein – gene is characterised as ribosomal protein (x); Pot. rel. info. content – potential relative information content calculated by new reduction heuristics (MARE); No. of taxa in data set – amount of taxa in the original data set; present in data subset – indicates the presence of that gene in the optimal data set after performing matrix reduction; No. of taxa in data subset – amount of taxa in the optimal data subset.

Gene ID	Protein ID	Gene / Description Ensembl Arch. 02/07 / InParanoid v6.1	Gene / Description InParanoid v6.1 (Aaeg) / HomoloGene (Hsap)	Other studies [shortcut]	Rib. Protein	Pot. rel. info. content	No. of taxa in data set	present in data subset	No. of taxa in data subset
12061	FBpp0078222	ADP-ribosylation factor 1				0.92	107	x	88
11924	FBpp0081153	Tubulin alpha-1 chain				0.92	149	x	102
11899	FBpp0078664	26S proteasome non-ATPase regulatory subunit 14				0.91	70	x	61
11735	FBpp0076890	26S protease regulatory subunit 8		ph, de, ba [nsf1-G]		0.90	81	x	76
11806	FBpp0081524	Beta-2 tubulin				0.90	127	x	96
11491	FBpp0083502	AP-2	clathrin coat assembly protein ap17			0.90	51	x	46
11394	FBpp0073292	Rpt3	26S protease regulatory subunit 6b Proteasome 26S subunit ATPase 4	de, ba [nsf1-L]		0.89	70	x	66
12024	FBpp0083645	AP-50, isoform A	clathrin coat associated protein ap-50			0.89	56	x	54
11846	FBpp0082140	Vacuolar ATP synthase subunit B		ph, de, ba [vatb]		0.89	74	x	69
11362	FBpp0084434	Histone H2A				0.88	80	x	70
11958	FBpp0078984	smt3				0.87	74	x	62
11637	FBpp0086701	40S ribosomal protein S23		ph, de, ba [rps23]	x	0.86	142	x	95
11460	FBpp0083906	26S protease regulatory subunit 4		ph, de, ba [nsf1-M]		0.86	74	x	69
11547	FBpp0085265	Elongation factor 2		ph, de, ba [ef2-EF2]		0.86	71	x	65
12071	FBpp0088250	ATP synthase beta chain, mitochondrial precursor				0.86	119	x	95
11624	FBpp0081592	AP-47	clathrin coat assembly protein ap-1			0.85	58	x	55
11511	FBpp0076145	CG6767-PB, isoform B	ribose-phosphate pyrophosphokinase 1			0.85	61	x	59
11609	FBpp0083843	Tat-binding protein-1	26S protease regulatory subunit 6a	ph, de, ba [nsf1-K]		0.85	78	x	71
11484	FBpp0088174	CG1970-PA	NADH-ubiquinone oxidoreductase fe-s protein 2 (ndufs2)			0.84	77	x	67
11902	FBpp0087084	GTP-binding protein 128up				0.84	58	x	55
11442	FBpp0077792	Splicing factor U2af 38 kDa subunit		de [u2snmp]		0.83	58	x	53
11552	FBpp0071808	60S ribosomal protein L23		ph, du, de, ba [rpl23a]	x	0.83	127	x	91
11760	FBpp0074520	Cdc42 homolog	rac GTPase			0.82	68	x	60
11645	FBpp0073446	Heat shock 70 kDa protein cognate 3 precursor		ph, de, ba [hsp70-E]		0.82	64	x	58
11868	FBpp0079999	Vacuolar ATP synthase catalytic subunit A isoform 2		ph, de, ba [vata]		0.82	51	x	48
11762	FBpp0078847	CG9140-PA	NADH-ubiquinone oxidoreductase flavoprotein 1 (ndufv1)			0.81	71	x	63
11377	FBpp0082724	SF2	arginine/serine-rich splicing factor			0.80	47	x	46
11759	FBpp0081401	CG8351-PA	chaperonin	ph, de, ba [cct-N]		0.78	64	x	60
11635	FBpp0080639	40S ribosomal protein S26		ph, de, ba [rps26]	x	0.78	124	x	93
11983	FBpp0082535	Tropomyosin-2				0.77	129	x	98
11617	FBpp0079992	CG5525-PA	chaperonin T-complex protein 1 subunit delta	ph, de, ba [cct-D]		0.77	73	x	67
11639	FBpp0085586	40S ribosomal protein S18		ph, du, de, ba [rps18]	x	0.77	136	x	97
11366	FBpp0077571	Enolase				0.77	94	x	78
11634	FBpp0083684	T-complex protein 1 subunit alpha	chaperonin	ph, de, ba [cct-A]		0.76	64	x	62
11393	FBpp0075700	Eukaryotic translation initiation factor 2 beta subunit		ph, de, ba [if2b]		0.76	70	x	60
11379	FBpp0071226	CG7033-PB, isoform B	chaperonin	ph, de, ba [cct-B]		0.76	66	x	63
11893	FBpp0072197	26S proteasome non-ATPase regulatory subunit 7				0.76	62	x	58
12097	FBpp0085919	Polyadenylate-binding protein				0.76	60	x	54
11514	FBpp0074180	40S ribosomal protein S5a		ph, ba [rps5]	x	0.75	148	x	108
11750	FBpp0073328	GTP-binding nuclear protein Ran				0.75	88	x	77
11681	FBpp0082459	CG3731-PB, isoform B	mitochondrial processing peptidase beta subunit	du [rpl27]		0.75	88	x	79

## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

11874	FBpp0079187	Guanine nucleotide-binding protein beta subunit-like protein		0.75	133	x	104
11962	FBpp0080495	Vacuolar ATP synthase subunit H		0.75	66	x	61
11965	FBpp0077142	60S ribosomal protein L27a	ph, de, ba [rpl27]	x 0.75	136	x	95
11450	FBpp0086603	Proteasome p44.5 subunit, isoform B		0.75	76	x	70
11917	FBpp0082464	VhaPPA1-1	vacuolar ATP synthase proteolipid subunit	0.74	71	x	66
11411	FBpp0082516	Heat shock 70 kDa protein cognate 4		0.74	109	x	92
11660	FBpp0078024	26S proteasome non-ATPase regulatory subunit 4		0.74	62	x	61
11385	FBpp0088565	Eukaryotic initiation factor 3 p66 subunit		0.74	72	x	66
11642	FBpp0077419	Phosphoglycerate kinase		0.74	79	x	72
11695	FBpp0077741	lesswright, isoform A	ubiquitin-conjugating enzyme E2 i	0.74	65	x	60
11587	FBpp0073626	40S ribosomal protein S15Aa	ph, de, ba [rps22a]	x 0.73	119	x	91
11829	FBpp0071794	ATP synthase alpha chain, mitochondrial precursor		0.73	103	x	92
12012	FBpp0074825	Catalase		0.73	63	x	61
12121	FBpp0086269	Ribosomal protein S15, isoform B	ph, du, de, ba [rps15]	x 0.73	133	x	97
11479	FBpp0081234	Probable small nuclear ribonucleoprotein Sm D2	du [small nuclear ribonucleoprotein polypeptide D2]	0.72	57	x	50
11798	FBpp0085483	Vacuolar ATP synthase 16 kDa proteolipid subunit		0.72	98	x	81
11848	FBpp0086468	Vacuolar ATP synthase subunit D 1		0.72	75	x	62
11627	FBpp0080691	Probable 26S proteasome non-ATPase regulatory subunit 3		0.72	65	x	62
11454	FBpp0073847	Adenosylhomocysteinase	ph, de [ Sadhchydrolase-E1]	0.72	94	x	81
12054	FBpp0077637	CG5001-PA	DNA-J/hsp40	0.72	65	x	60
11911	FBpp0078134	60S acidic ribosomal protein P0	ph, de, ba [rpp0]	x 0.72	148	x	108
12019	FBpp0076393	Isocitrate dehydrogenase, isoform F		0.71	79	x	66
11375	FBpp0077716	60S acidic ribosomal protein P1	du, de, ba [ria2-B]	x 0.71	134	x	89
11855	FBpp0082571	Surfeit locus protein 4 homolog		0.71	60	x	57
11583	FBpp0082788	T-complex protein 1 subunit gamma	ph, de, ba [cct-G]	0.71	57	x	55
11563	FBpp0081581	Calreticulin precursor		0.70	100	x	87
11429	FBpp0086381	CG8446-PA	lipoyltransferase 1	0.69	57	x	55
12033	FBpp0084585	CG5590-PA	short-chain dehydrogenase	0.69	72	x	65
11437	FBpp0078532	CG9769-PA	eukaryotic translation initiation factor 3f eif3f	0.69	73	x	62
11577	FBpp0082062	Proteasome subunit alpha type 2	ph, de, ba [psma-D]	0.69	64	x	54
11534	FBpp0071451	Proteasome subunit alpha type 4		0.69	76	x	64
11591	FBpp0072968	CG32278-PB, isoform B	stress-associated endoplasmic reticulum protein family member 2	0.68	104	x	76
11603	FBpp0077740	Signal peptide protease		0.68	58	x	57
11910	FBpp0072801	60S ribosomal protein L8	ph, du, de, ba [rpl2]	x 0.67	134	x	101
11802	Fbpp0071279	Oligosaccharyltransferase 48kD subunit	Dolichyl-diphosphooligosaccharide protein glycosyltransferase	0.67	68	x	64
11555	FBpp0080395	CaBP1	protein disulfide-isomerase A6 precursor	0.67	72	x	64
12120	FBpp0081780	Arginine methyltransferase 1		0.67	67	x	62
11814	Fbpp0076960	CG1532-PA	lactoylglutathione lyase	0.66	61	x	56
11567	FBpp0086066	Proteasome subunit alpha type 5	ph, de, ba [psma-A]	0.66	66	x	61
11451	FBpp0076152	40S ribosomal protein S9	ph, ba [rps9]	x 0.66	118	x	86
11710	FBpp0080724	Ribosomal protein L30, isoform A	ph, du, de, ba [rpl30]	x 0.66	117	x	88
11580	FBpp0110423	ribosomal protein L5	ph, de, ba [rpl5]	x 0.65	133	x	105
11976	FBpp0085489	Succinate dehydrogenase [ubiquinone] iron-sulfur protein, mitochondrial precursor		0.65	73	x	64
12029	FBpp0082985	CG7998-PA	malate dehydrogenase	0.65	86	x	75
11849	FBpp0072312	60S ribosomal protein L19	ph, du, de, ba [rpl19a]	x 0.65	134	x	93
12047	FBpp0084901	CG7834-PB, isoform B	electron transfer flavoprotein beta-subunit	0.65	73	x	68
11350	FBpp0076859	Uev1A, isoform B	ubiquitin-conjugating enzyme	0.64	62	x	58
11386	FBpp0079640	CG5362-PA	malate dehydrogenase	0.64	90	x	78
12112	FBpp0072250	Inorganic pyrophosphatase		0.63	71	x	59
11428	FBpp0073989	Proteasome subunit alpha type 7-1		0.63	72	x	65
11711	FBpp0075766	60S ribosomal protein L10a-2	ph, de, ba [rpl1]	x 0.63	133	x	103
11499	FBpp0070430	CG8636-PA	eukaryotic translation initiation factor eukaryotic translation initiation factor 3 subunit 4	du [eukaryotic translation initiation factor 3, subunit 4 delta]	0.63	81	x 71
11652	FBpp0085889	Eip55E	cystathionine beta-lyase	0.63	72	x	63
11378	FBpp0079472	yippee interacting protein 2		0.63	79	x	71
11919	FBpp0083371	40S ribosomal protein S20	ph, du, de, ba [rps20]	x 0.63	135	x	95
11772	FBpp0087186	walrus, isoform B	electron transport oxidoreductase	0.62	68	x	61
11928	FBpp0070047	60S ribosomal protein L10	ph, de, ba [grc5]	x 0.62	155	x	109
11932	FBpp0070871	Lethal (1), isoform A	citrate synthase	0.62	70	x	65
11380	FBpp0084617	60S ribosomal protein L4	ph, de, ba [rpl4B]	x 0.62	134	x	107

## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

11820	FBpp0076804	Thioredoxin-like			0.61	75	x	69	
11707	FBpp0088441	40S ribosomal protein S7	ph [rps7]	x	0.61	134	x	103	
11663	FBpp0088522	Ubiquitin conjugating enzyme 10			0.61	69	x	62	
11912	FBpp0074599	Clathrin light chain			0.61	79	x	68	
12046	FBpp0088505	Annexin-B9			0.61	76	x	64	
11992	FBpp0087164	Erp60, isoform B	protein disulfide isomerase		0.61	101	x	87	
11754	FBpp0071766	40S ribosomal protein S16	ph, du, de, ba [rps16]	x	0.60	138	x	101	
11984	FBpp0100039	Voltage-dependent anion-selective channel			0.60	104	x	84	
11773	FBpp0075382	Proteasome 2 subunit	ph, de, ba [psmb-K]		0.60	92	x	81	
11847	FBpp0088242	40S ribosomal protein S3a	ph, de, ba [rps1]	x	0.59	139	x	103	
11649	FBpp0076848	CG4769-PA	cytochrome C1		0.58	93	x	82	
12040	FBpp0084306	Ribosomal protein L27		x	0.58	129	x	87	
12035	FBpp0073344	Glutamine synthetase 2, cytoplasmic			0.57	90	x	79	
12115	FBpp0087972	cathD	cathepsin d		0.57	96	x	81	
12093	FBpp0082645	NADH:ubiquinone reductase 23kD subunit precursor			0.57	70	x	66	
12081	FBpp0084762	Elongation factor 1-gamma			0.56	137	x	109	
11619	FBpp0086103	60S ribosomal protein L18a	ph, du, de, ba [rpl20]	x	0.56	137	x	97	
11869	FBpp0077580	Rieske iron-sulfur protein, isoform B	ubiquinol-cytochrome c reductase iron-sulfur subunit	du [Ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide I] ph, ba [rps4]	0.56	97	x	80	
11391	FBpp0075618	40S ribosomal protein S4		x	0.55	144	x	105	
11584	FBpp0081488	Proteasome subunit beta type 3	ph, du, de, ba [psmb-I]		0.54	81	x	72	
11383	FBpp0086973	Nascent polypeptide-associated complex alpha subunit			0.53	94	x	81	
11778	FBpp0087608	60S ribosomal protein L31	ph, de, ba [rpl31]	x	0.53	122	x	89	
11844	FBpp0099686	40S ribosomal protein S8	ph, du, ba [rps8]	x	0.53	152	x	104	
11618	FBpp0085166	Ribosomal protein L6, isoform B	ph, de, ba [rpl6]	x	0.47	145	x	106	
11841	FBpp0110173	hydrogen-transporting ATP synthase, G-subunit, putative			0.46	110	x	78	
12122	FBpp0078354	60S ribosomal protein L13A		x	0.45	136	x	98	
12123	FBpp0083376	Ribosomal protein S30, isoform B	du [Ubiquitin-like FUBI and riboson	x	0.44	136	x	94	
12074	FBpp0072084	CG3195-PA, isoform A	60S ribosomal protein L12	ph, du, de, ba [rpl12b]	x	0.44	136	x	100
11793	FBpp0076602	Ribosomal protein L18	ph, du, de, ba [rpl18]	x	0.42	124	x	95	
11417	FBpp0087352	Ras-related protein Rab-3			0.88	34			
11816	Fbpp0079447	Pka-C1: cAMP-dependent protein kinase catalytic subunit	cAMP-dependent protein kinase catalytic subunit		0.87	44			
11387	FBpp0075260	diablo			0.86	36			
12009	FBpp0088695	CG2944-PF, isoform F	splA/ryanodine receptor domain and SOCS box containing 4		0.84	30			
11405	FBpp0077302	Protein mothers against dpp			0.83	29			
12073	FBpp0074756	reptin			0.83	46			
11755	FBpp0083248	CG10889-PA	zinc finger CCCH-type containing 12B		0.82	21			
11783	FBpp0079615	Transcription initiation factor IIB			0.81	46			
11860	FBpp0070208	SNF1A/AMP-activated protein kinase, isoform B			0.81	35			
11803	FBpp0079634	CG5343-PA	orf protein		0.81	47			
12118	FBpp0099616	cAMP-dependent protein kinase type I regulatory subunit			0.80	52			
11398	FBpp0088599	Potassium voltage-gated channel protein Shaker	voltage-gated potassium channel		0.80	19			
11954	FBpp0079565	Putative ATP-dependent RNA helicase me31b			0.80	43			
11509	FBpp0087094	Small nuclear ribonucleoprotein SM D3			0.79	43			
12087	FBpp0082743	COP9 signalosome complex subunit 5			0.79	46			
11865	FBpp0070361	Unc-76, isoform B			0.79	31			
11680	FBpp0088583	CG11266-PG, isoform G	splicing factor		0.78	49			
11989	FBpp0083135	CG5451-PA	WD-repeat protein		0.78	37			
11797	AAEL007662-PA	casein kinase			0.78	46			
12084	FBpp0079951	Ef1-like factor			0.78	28			
11850	FBpp0099884	UGP, isoform A			0.77	49			
11496	FBpp0078469	Katanin 60	de [nsf1-N]		0.77	34			
11687	FBpp0084528	CG5934-PA			0.76	36			
12070	AAEL005833-PA	cytosolic purine 5-nucleotidase			0.76	36			
11956	FBpp0086942	Guanine nucleotide-binding protein G(q) subunit alpha			0.76	36			
11616	FBpp0086375	Lissencephaly-1 homolog			0.76	46			
11673	FBpp0083973	Syntaxin-1A			0.75	38			
11991	FBpp0083588	CG6439-PA	isocitrate dehydrogenase		0.75	56			
12060	FBpp0074486	6-phosphofructo-2-kinase, isoform I			0.75	41			
11384	FBpp0081448	CG11990-PA	cdc73 domain protein		0.75	30			



## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

11542	FBpp0080659	Sterol carrier protein X-related thiolase		0.74	54
11763	FBpp0072052	Guanine nucleotide-binding protein G(s), alpha subunit		0.74	27
11700	FBpp0079629	RluA-1, isoform C		0.74	26
11589	FBpp0083573	Probable ATP-dependent RNA helicase pitchoune		0.74	38
11940	FBpp0085430	CG10465-PA	potassium channel tetramerisation domain containing 10	0.74	33
12065	FBpp0110163	CAMP-dependent protein kinase catalytic subunit		0.74	50
12007	FBpp0074691	tricornered		0.74	28
11864	FBpp0073387	DNA-directed RNA polymerase II largest subunit	du [DNA directed RNA polymerase II polypeptide C]	0.74	15
11921	FBpp0085737	Succinate dehydrogenase [ubiquinone] flavoprotein subunit, mitochondrial precursor		0.73	35
11406	FBpp0083112	endophilin A, isoform B		0.73	30
11726	FBpp0088499	Protein ariadne-1		0.73	42
11640	FBpp0071553	CG4279-PA	Sm protein G putative	0.73	42
11564	FBpp0085131	CG31005-PA	trans-prenyltransferase	0.73	32
11508	FBpp0080261	Suppressor of hairless protein		0.73	18
11788	FBpp0110435	synaptosomal associated protein		0.73	38
11672	FBpp0071424	Inosine-5'-monophosphate dehydrogenase		0.73	46
11610	FBpp0070859	Spliceosomal protein on the X		0.73	28
11980	FBpp0085902	GTP-binding-protein		0.72	46
11523	FBpp0078624	CG14641-PA	RNA binding motif protein	0.72	37
11990	FBpp0081290	ADP-ribosylation factor-like protein 8		0.72	49
11768	FBpp0074278	CG6842-PA	skd/vacuolar sorting	0.72	43
11934	FBpp0070250	CG32810-PB	potassium channel tetramerisation domain containing 5	0.72	34
11578	FBpp0071600	Rae1		0.72	46
11436	FBpp0084036	atlastin, isoform B		0.72	41
11821	FBpp0070883	Serine/threonine-protein phosphatase PP-V	de [stcproptase2a-c]	0.72	47
11490	FBpp0081483	Aryl hydrocarbon receptor nuclear translocator homolog		0.72	20
11796	FBpp0081704	pontin		0.71	34
11549	FBpp0076078	Ard1, isoform A		0.70	51
12038	FBpp0082507	CG4203-PA	KIAA0892	0.70	26
12088	FBpp0079676	Stress-activated protein kinase JNK		0.70	26
11718	FBpp0086599	CG32105-PB	LIM homeobox transcription factor 1, alpha	0.70	20
11785	FBpp0080801	Tyrosine-protein phosphatase Lar precursor		0.70	17
11572	FBpp0086790	Elongation factor Tu mitochondrial		0.70	61
12068	FBpp0082129	Malic enzyme, isoform A		0.70	47
11402	FBpp0070794	Males-absent on the first protein		0.70	28
11535	FBpp0073872	CG9281-PC, isoform C	ATP-dependent transporter	0.70	38
11536	FBpp0083954	CG31137-PE, isoform E	carbon catabolite repressor protein	0.70	24
11419	FBpp0081437	CG11963-PA	succinyl-coa synthetase beta chain	0.69	57
11859	FBpp0074609	Soluble NSF attachment protein		0.69	52
11641	FBpp0075372	Echinoderm microtubule-associated protein-like CG13466	WD-repeat protein	0.69	16
11815	FBpp0078880	Cpr: NADPH-cytochrome P450 reductase	NADPH cytochrome P450	0.68	53
11573	FBpp0073010	Succinyl-CoA ligase [GDP-forming] alpha-chain, mitochondrial precursor	ph, de, ba [suca]	0.68	62
12059	FBpp0078764	CG7236-PA	cdk11/4	0.68	17
11838	FBpp0083687	26S proteasome non-ATPase regulatory subunit 6		0.68	63
12108	FBpp0086605	CG12858-PA	major facilitator superfamily domain containing 6	0.68	20
12082	FBpp0073090	Transcription factor IIE		0.68	36
12002	FBpp0081336	steamer duck, isoform C		0.68	39
11733	FBpp0087535	CG1513-PA	oxysterol binding protein 9	0.68	30
12105	FBpp0083549	CG6560-PA	ADP-ribosylation factor arf	0.67	36
11935	FBpp0087870	Protein peanut		0.67	35
11403	FBpp0077414	Congested-like trachea protein		0.67	45
11418	FBpp0080281	Transcription elongation factor S-II	transcription elongation factor s-ii	0.67	51
11480	FBpp0073003	eIF5B		0.67	23
11988	FBpp0077735	Notchless		0.67	38
12010	FBpp0081216	Transcription initiation factor IIF alpha subunit		0.67	43
11823	FBpp0077996	Rab26		0.67	27
11824	FBpp0076647	UDP-glucose 6-dehydrogenase		0.67	32
11482	FBpp0077720	CG4164-PA	DNA-J/hsp40	0.67	48
11787	FBpp0072621	phosphocholine cytidyltransferase 1, isoform D		0.67	35

## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

11604	FBpp0072122	Calcium-transporting ATPase sarcoplasmic/endoplasmic reticulum type		0.67	25
11392	FBpp0088988	Glutamate dehydrogenase, mitochondrial precursor		0.67	57
11628	FBpp0075684	Probable small nuclear ribonucleoprotein Sm D1		0.67	54
11805	FBpp0088775	CG33096-PB, isoform B	family with sequence similarity 108, member C1	0.67	29
11495	FBpp0077171	CG3714-PB, isoform B	nicotinate phosphoribosyltransferase	0.66	28
11854	FBpp0073119	Protein ROP		0.66	36
11993	FBpp0081350	tex		0.66	41
11808	FBpp0082867	Guanine nucleotide-binding protein-like 3 homolog	GTP-binding protein-invertebrate	0.66	56
11545	FBpp0070808	CG32758-PA	sorting nexin	0.66	26
12003	FBpp0074543	Tao-1, isoform E		0.66	30
11857	FBpp0085619	Proliferating cell nuclear antigen	du [Proliferating cell nuclear antigen]	0.66	58
11502	FBpp0073600	CG1640-PA, isoform A	alanine aminotransferase	0.66	63
11757	FBpp0072672	Spectrin alpha chain		0.66	24
11916	FBpp0078400	Splicing factor 3A subunit 3		0.65	48
11507	FBpp0081840	CG17184-PB, isoform B	ADP-ribosylation factor interacting protein 2	0.65	27
11588	FBpp0087472	CG12140-PA	electron transfer flavoprotein-ubiquinone oxidoreductase	0.65	34
11728	FBpp0099695	Dystrobrevin-like, isoform A		0.65	19
11716	FBpp0071992	no extended memory, isoform B		0.65	31
12018	FBpp0083611	Pyruvate kinase		0.65	77
11830	FBpp0087865	Rs1		0.65	29
12076	FBpp0078099	CG7145-PD, isoform D	pyrroline-5-carboxylate dehydrogenase	0.65	57
11556	FBpp0079843	CG14939-PA	cyclin Y	0.65	31
11553	FBpp0071285	Puff-specific protein Bx42		0.65	43
11955	FBpp0110314	conserved hypothetical protein		0.65	36
11667	FBpp0077214	CG17593-PA	coiled-coil domain containing 47	0.65	49
12069	FBpp0071046	Protein bys		0.65	40
11709	FBpp0074022	CG9911-PA, isoform A	endoplasmic reticulum resident protein (ERp44) putative	0.65	44
11389	FBpp0081988	Putative inner dynein arm light chain	axonemal inner arm dynein light chain	0.65	24
12051	FBpp0072144	Probable eukaryotic translation initiation factor 6	ph, de, ba [t6]	0.65	60
11525	FBpp0086098	eIF3-S9, isoform B		0.64	76
11432	FBpp0079642	CG33303-PA	ribophorin	0.64	68
11712	FBpp0070249	CG14782-PA	pleckstrin homology domain containing, family F (with FYVE domain) member 2	0.64	31
11905	FBpp0078433	DNA-directed RNA polymerases I, II, and III 14.4 kDa polypeptide		0.64	47
11598	FBpp0075202	CG5284-PA, isoform A	chloride channel protein 3	0.64	26
11853	FBpp0081617	CG8500-PA	MRAS2 putative	0.64	19
11890	FBpp0086340	mrj, isoform D		0.64	49
11801	FBpp0072419	Tudor-SN	ebna2 binding protein P100	0.64	55
11455	FBpp0082728	belphégor		0.64	34
12057	FBpp0076921	lethal (1) G0269		0.64	25
11574	FBpp0077676	Clipper		0.64	30
11971	FBpp0070873	Transmembrane GTPase Marf		0.64	39
11891	FBpp0081958	CG18347-PA	mitochondrial glutamate carrier protein	0.64	32
11786	FBpp0084191	CG11859-PA	serine/threonine-protein kinase rio2 (rio kinase 2)	0.64	35
11629	FBpp0079617	CHIP		0.63	46
11632	FBpp0078997	nop5		0.63	49
11608	FBpp0078606	ATP-dependent RNA helicase abstrakt		0.63	31
11351	FBpp0070651	cap binding protein 80, isoform A		0.63	28
11975	FBpp0080282	crinkled, isoform A		0.63	17
11349	FBpp0083972	4EHP	eukaryotic translation initiation factor 4e type	0.63	56
11529	FBpp0076244	Probable signal recognition particle 68 kDa protein	srp68	0.63	50
11355	FBpp0081374	belle	DEAD box ATP-dependent RNA helicase	0.63	41
12053	FBpp0072788	CG9018-PB, isoform B	regulation of nuclear pre-mRNA domain containing 1B	0.63	38
11368	FBpp0075729	RhoGAP68F		0.63	36
11831	FBpp0078191	CG6838-PB, isoform B	ADP-ribosylation factor GTPase activating protein 2	0.63	55
11613	FBpp0086590	CG12797-PA	WD-repeat protein	0.63	42
11799	FBpp0078371	MLF1-adaptor molecule		0.63	30
11883	FBpp0089034	Armadillo segment polarity protein		0.63	21
11880	FBpp0071407	Mannosyl-oligosaccharide alpha-1,2-mannosidase isoform 2		0.63	30
11771	FBpp0078161	Tenascin major		0.63	16

## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

11843	FBpp0078887	CG9523-PA	<i>FIC domain containing</i>	0.63	28
11531	FBpp0085140	CG31004-PB, isoform B	<i>sushi domain containing 2</i>	0.63	23
11571	AAEL012316-PA	arsenical pump-driving ATPase		0.63	31
12031	AAEL014285-PA	growth hormone inducible transmembrane protein	du [growth hormone inducible transmembrane protein]	0.63	59
11501	FBpp0074151	Probable small nuclear ribonucleoprotein G	du [small nuclear ribonucleoprotein polypeptide G]	0.63	61
11520	FBpp0073134	Fumarylacetoacetase		0.62	52
12042	FBpp0062569	CG6194-PA	<i>ATG4 autophagy related 4 homolog D</i>	0.62	29
11647	FBpp0078811	Tetraspanin 26A		0.62	34
11964	FBpp0089047	Voltage-dependent calcium channel type D alpha-1 subunit		0.62	13
11739	FBpp0087699	Receptor mediated endocytosis 8		0.62	17
12049	FBpp0100147	conserved membrane protein at 44E, isoform A		0.62	29
11742	FBpp0070418	CG16903-PA	cyclin I	0.62	38
11822	FBpp0078893	CG9547-PA	acyl-CoA dehydrogenase	0.62	51
11424	FBpp0079870	escl, isoform A		0.62	40
11413	FBpp0086223	Flap endonuclease 1		0.62	39
11705	FBpp0075485	Protein frizzled precursor		0.62	25
11737	FBpp0087722	Dynamitin		0.62	53
11701	AAEL010002-PB	5-formyltetrahydrofolate cyclo-ligase		0.62	53
11939	FBpp0076523	Protein henna		0.62	59
11576	FBpp0070104	Beta-amyloid-like protein precursor		0.62	37
11527	FBpp0084894	CG31033-PB, isoform B	<i>ATG16 autophagy related 16-like 1</i>	0.62	18
12102	FBpp0081633	CG9461-PA	F-box only protein	0.62	23
12092	FBpp0088153	Eph receptor tyrosine kinase, isoform D		0.62	18
11631	FBpp0070368	6-phosphogluconate dehydrogenase, decarboxylating		0.62	56
12075	FBpp0084499	CG6051-PA	lateral signaling target protein	0.62	29
11903	FBpp0072723	CG1140-PA, isoform A	succinyl-coa:3-ketoacid-coenzyme a transferase	0.61	43
11896	FBpp0086289	HMG Coenzyme A synthase, isoform A		0.61	42
11922	FBpp0079976	PICK1, isoform B		0.61	30
11996	FBpp0078360	sec23, isoform B		0.61	24
11625	FBpp0088955	Protein tumorous imaginal discs, mitochondrial precursor		0.61	56
12066	FBpp0070793	CG3016-PA	ubiquitin-specific protease	0.61	23
11913	FBpp0083976	Rox8, isoform F		0.61	39
12062	FBpp0088862	Hypothetical protein CG7816		0.61	36
11416	FBpp0088910	CG1732-PB, isoform B	sodium/chloride dependent neurotransmitter transporter	0.61	30
11530	FBpp0070469	Hypothetical protein CG32795 in chromosome 1		0.61	44
11565	FBpp0077998	CG7338-PA	ribosome biogenesis protein tsr1	0.61	50
11881	FBpp0082624	CG4525-PA	<i>tetratricopeptide repeat domain 26</i>	0.61	21
12103	FBpp0084774	CG1458-PA	<i>CDGSH iron sulfur domain 2</i>	0.61	68
12014	FBpp0075942	CG7628-PA	phosphate transporter	0.61	29
11607	FBpp0071259	CG12135-PA	<i>CWC15 spliceosome-associated protein homolog</i>	0.61	51
11930	FBpp0074330	CG6179-PA	<i>nitric oxide synthase interacting protein</i>	0.61	44
12016	FBpp0070751	RhoGAP5A, isoform A		0.61	24
11630	FBpp0079643	CG5366-PA	cullin-associated NEBD8-dissociated protein 1	0.60	20
12013	FBpp0087648	RNA-binding protein 8A		0.60	56
11666	FBpp0072660	Hsp90 co-chaperone Cdc37		0.60	56
11483	FBpp0077886	Lipoic acid synthase, isoform B	lipoic acid synthetase	0.60	48
11390	FBpp0075731	Neurexin-4 precursor		0.60	17
11929	FBpp0074822	Aut1		0.60	40
11396	FBpp0070064	Molybdenum cofactor synthesis protein cinnamon	molybdopterin biosynthesis protein	0.60	29
11719	FBpp0081331	CG10153-PA	<i>trafficking protein particle complex 5</i>	0.60	35
11978	FBpp0087366	CG11777-PA	cyclophilin-10	0.60	30
11727	FBpp0071303	CG3004-PA	vegetalible incompatibility protein HET-E-1 putative	0.60	39
11600	FBpp0087340	CG7686-PA	<i>LTV1 homolog</i>	0.60	53
12107	FBpp0085500	CG3358-PB, isoform B	<i>TatD DNase domain containing 1</i>	0.60	33
11372	FBpp0087938	Nup44A, isoform A		0.60	40
11792	FBpp0071262	CG17446-PA	cpg binding protein	0.60	27
11840	FBpp0078381	CG2185-PA	calcineurin b subunit	0.60	61
11926	AAEL002852-PA	conserved hypothetical protein		0.60	16
11357	FBpp0074246	CG8142-PA	replication factor C 37-kDa subunit putative	0.60	42

## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

12011	FBpp0070162	CG11642-PC, isoform C	translocation associated membrane protein		0.60	66
11447	FBpp0086703	CG8394-PA	amino acid transporter		0.60	21
11512	FBpp0074937	NUCB1			0.60	51
11671	FBpp0071392	CG32687-PA	internalin A putative		0.60	46
11606	FBpp0077047	lethal (1) G0196, isoform E			0.60	21
12094	FBpp0079675	CG5676-PA			0.59	51
11518	FBpp0073557	CG4332-PA	CLPTM1-like		0.59	35
11925	FBpp0071138	Probable phenylalanyl-tRNA synthetase alpha chain			0.59	37
12079	FBpp0078891	CG9543-PA	coatomer protein complex, subunit epsilon		0.59	58
11654	FBpp0077263	Probable tyrosyl-DNA phosphodiesterase			0.59	33
11407	FBpp0076124	Ubiquitin-conjugating enzyme E2-22 kDa	ubiquitin-conjugating enzyme E2-25kDa		0.59	50
11643	FBpp0071688	Protein ariadne-2			0.59	33
11731	FBpp0085222	lethal (3) s1921			0.59	47
12056	FBpp0081800	Sorbitol dehydrogenase-2			0.59	69
11767	FBpp0086875	F-box/SPRY-domain protein 1			0.59	21
11944	FBpp0071269	CG12121-PA	lung seven transmembrane receptor		0.59	26
11561	FBpp0072481	CG13887-PB, isoform B	B-cell receptor-associated protein bap	du [B-cell receptor-associated protein 31]	0.59	70
11997	FBpp0079914	Threonyl-tRNA synthetase, isoform C		ba [trs]	0.59	27
11834	FBpp0077129	CG15433-PA	elongator component putative		0.59	35
11538	FBpp0087714	CG8080-PA	chromosome 5 open reading frame 33		0.59	32
11478	FBpp0074792	CG6812-PA	sideroflexin 123		0.59	33
12000	FBpp0073354	CG1749-PA	ubiquitin-activating enzyme E1		0.59	46
11694	FBpp0070933	Serine/threonine-protein kinase			0.59	16
11817	FBpp0110272	multiple C2 domain and transmembrane region protein			0.59	20
11570	FBpp0071229	CG7039-PA	ARL3 putative		0.59	42
11692	FBpp0079812	Replication factor C 38kD subunit			0.59	42
11761	FBpp0088881	supercoiling factor, isoform B			0.59	49
11524	FBpp0086373	CysteinyI-tRNA synthetase			0.59	38
12110	FBpp0099935	CG11919-PA, isoform A	peroxisome assembly factor-2 (peroxisomal-type ATPase 1)		0.59	26
11884	FBpp0071478	CDK5RAP3-like protein			0.59	44
11427	FBpp0075042	rogdi, isoform A			0.59	29
11960	FBpp0087073	CG8841-PC, isoform C	chromosome 17 open reading frame 28		0.59	18
11488	FBpp0079946	Probable ribosome production factor 1	U3 small nucleolar ribonucleoprotein protein imp4		0.59	44
12080	FBpp0082525	CG4338-PA	chromosome 16 open reading frame 42		0.59	37
11358	FBpp0078721	thickveins, isoform D			0.59	31
11677	FBpp0071189	CG12125-PA	family with sequence similarity 73, member B		0.59	24
12109	FBpp0075866	CG11660-PA, isoform A	serine/threonine-protein kinase rio1 (rio kinase 1)		0.58	33
12050	FBpp0081087	CG2656-PA	GPN-loop GTPase 3		0.58	41
11356	FBpp0080407	CG5861-PA	transmembrane protein 147		0.58	47
11651	FBpp0075139	CG4933-PA	o-sialoglycoprotein endopeptidase		0.58	29
11752	FBpp0088329	Calcium-dependent secretion activator			0.58	13
11878	FBpp0071669	GlcT-1			0.58	28
11657	FBpp0070443	40S ribosomal protein S12, mitochondrial precursor		x	0.58	38
11871	FBpp0074990	UDP-sugar transporter UST74c			0.58	29
11920	FBpp0074662	Rpn1			0.58	30
11953	FBpp0084464	BM-40-SPARC			0.58	68
11863	FBpp0083098	Mekk1, isoform B			0.58	17
11614	FBpp0099673	Tousled-like kinase, isoform D			0.58	24
11438	FBpp0078382	MTA1-like, isoform B			0.58	17
11381	FBpp0081659	lethal (3) IX-14			0.58	20
11686	FBpp0072142	Protein within the bgcn gene intron			0.57	38
11875	FBpp0084118	CG5805-PA	mitochondrial glutamate carrier putative		0.57	17
11360	FBpp0088059	Dpld (Protein dappled)			0.57	17
12090	FBpp0075734	CG6910-PA	myoinositol oxygenase		0.57	53
11734	AAEL010797-PA	RNA polymerase II holoenzyme component			0.57	31
11477	FBpp0078633	CG3756-PA	DNA-directed RNA polymerase		0.57	44
12048	FBpp0086107	Anaphase-promoting complex subunit 10			0.57	36
11987	FBpp0081601	CG9373-PA	myelinprotein expression factor		0.57	43
11882	FBpp0073588	CG1622-PA	PRP38 pre-mRNA processing factor 38		0.57	33

## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

11810	FBpp0086757	CG12295-PB: straightjacket	dihydropyridine-sensitive l-type calcium channel		0.57	16	
11668	FBpp0070389	mitochondrial ribosomal protein L14		x	0.57	38	
11858	FBpp0086063	Ngp			0.56	34	
11795	FBpp0084691	CG1646-PC, isoform C	PRP39 pre-mRNA processing factor 39 homolog		0.56	43	
11544	FBpp0085838	GDI interacting protein 3, isoform C			0.56	33	
11400	FBpp0071061	Integrin beta-PS precursor	integrin beta subunit		0.56	45	
11550	FBpp0076134	ATP synthase B chain, mitochondrial precursor			0.56	16	
11725	FBpp0076486	pebble, isoform D			0.56	18	
11804	FBpp0087806	CG8635-PA	zinc finger CCCH-type containing 15		0.56	45	
11674	FBpp0074121	CG9099-PA	density-regulated protein		0.56	54	
11839	FBpp0081520	Probable maleylacetoacetate isomerase 2			0.56	43	
11559	FBpp0110208	calnexin			0.56	59	
11794	FBpp0084559	rapynoid			0.56	19	
11656	FBpp0075168	Tyrosyl-tRNA synthetase			0.56	45	
11489	FBpp0071445	CG9236-PA	calcium and integrin-binding protein 1		0.56	22	
12111	FBpp0084144	CG11920-PA	U3 small nucleolar ribonucleoprotein protein imp4		0.56	45	
11774	FBpp0085422	O-glycosyltransferase, isoform B			0.56	19	
11683	FBpp0086129	Fat-spondin, isoform B			0.56	52	
11461	FBpp0081481	Protein neuralized			0.55	29	
11835	FBpp0079780	CG6724-PA	WD-repeat protein		0.55	45	
11974	FBpp0083581	CG6015-PA	pre-mRNA splicing factor prp17		0.55	35	
11354	FBpp0081552	CG8286-PA	tetratricopeptide repeat protein putative		0.55	48	
12063	FBpp0078685	Probable GDP-mannose 4,6 dehydratase			0.55	38	
12078	FBpp0087709	Mystery 45A			0.55	34	
11972	FBpp0072382	mrityu, isoform C			0.55	24	
11828	FBpp0082895	CG5840-PB, isoform B	pyroline-5-carboxylate reductase		0.55	51	
11866	FBpp0083076	Probable 28 kDa Golgi SNARE protein			0.55	37	
11382	FBpp0082888	Sur-8, isoform A			0.55	33	
11704	FBpp0081370	CG8036-PD, isoform D	transketolase I		0.55	41	
11487	FBpp0083131	Prp18			0.55	38	
11691	FBpp0071063	Glutamate--cysteine ligase			0.55	30	
11458	FBpp0074562	CG32528-PA	parvin		0.54	48	
11861	FBpp0074026	Katanin 80, isoform B			0.54	20	
11586	FBpp0074835	CG6841-PA	pre-mRNA splicing factor		0.54	25	
11959	FBpp0080906	La protein homolog			0.54	55	
11404	FBpp0080622	CG10333-PA	DEAD box ATP-dependent RNA helicase		0.54	20	
11467	FBpp0072979	CG11537-PB, isoform B	hippocampus abundant transcript 1		0.54	20	
11364	FBpp0075238	PDCD-5	programmed cell death 5	du, de [pace6]	0.54	53	
11370	FBpp0081276	pyd3			0.54	49	
11722	FBpp0080048	Coatomer subunit beta'			0.54	21	
11425	FBpp0076861	Kinesin-like protein at 64D			0.54	27	
11675	FBpp0076332	CG7112-PA	rab6 GTPase activating protein gapcena (rabgap1 protein)		0.54	23	
11895	FBpp0081475	CG18005-PA	red protein (ik factor) (cytokine ik)		0.54	30	
11827	FBpp0081719	Sirt6			0.54	29	
12045	FBpp0079832	CG6509-PB, isoform B	discs large protein		0.54	14	
11615	FBpp0070367	CG3835-PA, isoform A	D-lactate dehydrognease 2		0.54	31	
11937	FBpp0079577	Ubiquitin thioesterase otubain-like protein			0.54	41	
11546	FBpp0072404	mitochondrial ribosomal protein L17		ph, du, de, ba [rpl17]	x	0.54	38
12099	FBpp0085155	Coatomer protein, isoform B			0.54	28	
11915	FBpp0073739	MRNA-capping-enzyme			0.54	46	
11579	FBpp0077551	CG31938-PA	exosome component 3		0.54	31	
11776	FBpp0072455	Probable UDP-glucose 4-epimerase			0.54	47	
11446	FBpp0084351	CG6095-PB, isoform B	exocyst complex-subunit protein 84kDa-subunit putative		0.54	31	
11434	FBpp0074582	CG14232-PA	acyl-Coenzyme A binding domain containing 3		0.54	29	
11809	FBpp0085181	CG1800-PA: partner of drosha	double-stranded binding protein putative		0.53	23	
11521	FBpp0071095	CG10932-PA	acetyl-coa acetyltransferase mitochondrial		0.53	65	
12106	FBpp0075693	Probable phosphomannomutase			0.53	47	
12098	FBpp0086667	CG8531-PA	DnaJ (Hsp40) homolog, subfamily C, member 11		0.53	35	
11646	FBpp0075111	COP, isoform B			0.53	62	

## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

11426	FBpp0083921	CG5991-PC, isoform C		0.53	37	
11764	FBpp0073806	CG14407-PA	glutaredoxin	0.53	68	
12067	FBpp0088040	CG11107-PA	ATP-dependent RNA helicase	0.53	22	
11528	FBpp0080117	CG16865-PA	chromosome X open reading frame 56	0.53	34	
11769	FBpp0073649	CG11134-PA	APAF1 interacting protein	0.53	48	
11590	FBpp0085763	Exostosin-3		0.53	24	
12083	FBpp0075947	Multidrug-Resistance like Protein 1, isoform B		0.53	26	
11894	FBpp0071256	C12.2		0.53	27	
11409	FBpp0074844	CG3961-PB, isoform B	long-chain-fatty-acid coa ligase	0.53	35	
11441	FBpp0073635	CG11178-PB, isoform B	AVL9 homolog	0.53	25	
11456	FBpp0072830	misshapen, isoform E		0.53	31	
11444	FBpp0071232	AP-1, isoform E		0.53	17	
11423	FBpp0078070	CG9391-PA, isoform A	myo inositol monophosphatase	0.53	52	
11526	FBpp0074564	CG12703-PA	peroxisomal membrane protein 70 abcd3	0.52	23	
11708	FBpp0081576	eclair		0.52	65	
12043	FBpp0074736	CG8798-PA, isoform A	ATP-dependent Lon protease putative	0.52	27	
11745	FBpp0090943	CG33505-PA	WD-repeat protein	0.52	36	
12096	FBpp0087342	CG12343-PA	SYF2 homolog, RNA splicing factor	0.52	51	
12026	FBpp0084813	CG1907-PA	solute carrier family 25 (mitochondrial carrier; oxoglutarate carrier), member 11	0.52	49	
12116	FBpp0078694	mitochondrial ribosomal protein L24		x 0.52	50	
12114	FBpp0075560	CG10711-PA	conserved hypothetical protein	0.52	46	
11596	FBpp0076073	nudE		0.52	38	
11698	FBpp0078992	Gas41		0.52	31	
11904	AAEL010402-PA	DEAD box ATP-dependent RNA helicase		0.52	20	
11970	FBpp0083436	Exocyst complex component 6		0.52	36	
11756	FBpp0086641	Lamin-C		0.52	37	
11724	FBpp0081283	CG10903-PA	Williams Beuren syndrome chromosome region 22	0.52	42	
11721	AAEL004763-PA	conserved hypothetical protein		0.52	26	
11592	FBpp0070806	Lethal (1), isoform A		0.52	21	
11601	FBpp0081834	CG5214-PA	dihydrolipoamide succinyltransferase component of 2-oxoglutarate dehydrogenase	0.52	45	
11431	FBpp0074226	CG5703-PA	NADH-ubiquinone oxidoreductase 24 kDa subunit	du [NADH dehydrogenase (ubiquinone) flavoprotein 2, 24kDa]	0.52	71
11811	FBpp0079495	CG5885-PA	translocon-associated protein gamma subunit		0.52	84
11775	FBpp0074517	Glucose-6-phosphate 1-dehydrogenase		0.52	42	
11506	FBpp0082642	CG4225-PA	ABC transporter Mitochondrial ABC transporter 3	0.51	17	
11826	FBpp0110402	eukaryotic translation initiation factor 3, theta subunit		0.51	24	
11422	FBpp0085952	Dgp-1, isoform A		0.51	33	
11782	FBpp0073828	CG6227-PA	DEAD box ATP-dependent RNA helicase	0.51	17	
11685	FBpp0071217	Polycomb protein l(1)G0020		0.51	30	
11723	FBpp0081799	CG6465-PA	aminoacylase putative	0.51	52	
11376	FBpp0075938	NEDD8-activating enzyme E1 regulatory subunit	app binding protein	0.51	39	
11699	FBpp0075034	CG7728-PA	ribosome biogenesis protein	0.51	29	
11789	FBpp0084190	CG11858-PA	peptidyl-prolyl cis/trans isomerase putative	du [protein (peptidyl-prolyl cis/trans isomerase) NIMA-interacting 1]	0.51	40
11463	FBpp0087926	drosha		0.51	15	
11867	FBpp0074734	CG8793-PA	KIAA1012		0.50	17
11697	AAEL013319-PA	conserved hypothetical protein		ph, de [stbproptase2a-b]	0.50	15
11688	FBpp0075069	CG4169-PA	ubiquinol-cytochrome c reductase complex core protein		0.50	82
11952	FBpp0076782	Regulator of chromosome condensation		0.50	30	
11740	FBpp0074366	Histidyl-tRNA synthetase, isoform B		0.50	45	
11898	FBpp0087506	6-phosphofructokinase		0.50	30	
11892	FBpp0083899	Bifunctional aminoacyl-tRNA synthetase		0.50	24	
11473	FBpp0084489	DNA polymerase alpha subunit B		0.50	29	
12008	FBpp0078184	Secretory Pathway Calcium atpase, isoform C		0.50	17	
11513	FBpp0072531	CG9119-PA	chromosome 11 open reading frame 54	0.50	44	
11669	FBpp0073875	CG9245-PB, isoform B	phosphatidylinositol synthase	0.50	46	
11813	Fbpp0089153	smallmined CG8571-PB, isoform B	peroxisome assembly factor-2 (peroxisomal-type ATPase 1)	de [nsf2-B]	0.50	32
12023	FBpp0083842	3-hydroxy-3-methylglutaryl-coenzyme A reductase		0.50	35	
11659	FBpp0087353	CG16728-PA	G protein-coupled receptor kinase interacting ArfGAP 2	0.50	25	
11886	FBpp0080045	Two A-associated protein of 42kDa		0.50	37	
11852	FBpp0084032	CG6643-PB, isoform B	synaptotagmin putative	0.50	33	

## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

11982	FBpp0072779	CG1317-PB	ssm4 protein	0.50	25
11457	FBpp0082284	faiafeI, isoform C		0.50	17
11412	FBpp0070643	CG3564-PA	copii-coated vesicle membrane protein P24	0.50	72
11837	FBpp0074510	CG14211-PB	dual-specificity protein phosphatase putative	0.49	30
11449	FBpp0074285	3-hydroxyacyl-CoA dehydrogenase type-2	hydroxyacyl dehydrogenase	0.49	68
11653	FBpp0085609	Mediator complex subunit 8		0.49	38
11941	FBpp0085630	CG11208-PA	2-hydroxyphytanoyl-coa lyase	0.49	42
11914	FBpp0072495	CG13900-PB, isoform B	spliceosomal protein sap	0.49	23
11889	FBpp0099977	CG1410-PA, isoform A	GTP-binding protein lepa	0.49	30
11539	FBpp0074936	CG5589-PA	DEAD box ATP-dependent RNA helicase	0.49	36
11471	FBpp0080509	Aminopeptidase P		0.49	52
11730	FBpp0086954	Chromatin remodelling complex ATPase chain iswi		0.49	20
11459	FBpp0080319	lethal (2) 35Df		0.49	21
11909	FBpp0078319	CG2051-PC, isoform C	histone acetyltransferase type b catalytic subunit	0.49	38
11825	FBpp0070181	CG3704-PA	xpa-binding protein 1 (mbdin)	0.49	34
12039	FBpp0085873	Late endosomal/lysosomal Mp1-interacting protein homolog		0.49	51
11936	FBpp0086402	CG8386-PA	ubiquitin-fold modifier conjugating enzyme 1	0.49	53
12037	FBpp0080062	Skf6	du, de [rrp46-B]	0.49	37
11918	FBpp0087402	Caf1-105		0.49	29
11636	FBpp0084013	Golgin-84		0.49	25
11500	FBpp0088794	CG33298-PB, isoform B	phospholipid-transporting ATPase 1 (aminophospholipid flippase 1)	0.48	20
11566	FBpp0082288	neither inactivation nor afterpotential B		0.48	26
11784	FBpp0072058	Alpha-catenin-related, isoform B		0.48	21
12077	FBpp0077208	Exocyst complex component 2		0.48	33
11363	FBpp0083319		Signal recognition particle 72 kDa protein	0.48	50
11465	AAEL000324-PA	CG5589-PA, isoform B tyrosine-protein kinase dri		0.48	21
12017	FBpp0087867	Mih1		0.48	26
11741	FBpp0079735	Vacuolar protein sorting protein 72 homolog		0.48	31
11621	FBpp0072711	CG12091-PA	protein phosphatase 2c	0.48	48
11408	FBpp0082066	Interleukin enhancer-binding factor 2 homolog	interleukin enhancer binding factor	0.48	54
11720	AAEL002870-PA	Dipeptidyl-peptidase 3		0.48	55
12101	FBpp0076771	CG10467-PA	aldose-1-epimerase	0.48	42
11676	FBpp0075209	Signal sequence receptor	du [signal sequence receptor, beta precursor]	0.48	92
11644	FBpp0075535	Ral guanine nucleotide exchange factor 2, isoform A		0.48	21
11352	FBpp0075513	Hsc70Cb, isoform C		0.48	50
11714	FBpp0089006	CG32626-PD, isoform D	AMP deaminase	0.48	22
11833	FBpp0075151	multi-protein bridging factor, isoform B	du [endothelial-differentiation-related factor 1 isoform alpha]	0.48	74
11515	FBpp0083989	Dis3		0.48	32
11998	FBpp0078512	CG1126-PA	Bardet-Biedl syndrome 5	0.48	17
11435	FBpp0086042	CG6401-PA	glycosyltransferase	0.48	28
11790	FBpp0072468	CG6905-PA	cell division control protein	0.48	19
11558	FBpp0071277	Zpr1		0.48	43
11862	FBpp0077203	CG31957-PA	translation initiation factor 1A putative	0.48	34
12085	FBpp0086957	CG8632-PB, isoform B	solute carrier family 30 (zinc transporter), member 9	0.48	27
11469	FBpp0083440	Uridine 5'-monophosphate synthase	orotidine-5'-phosphate decarboxylase, putative	0.48	46
11474	FBpp0070180	CG3703-PA	RUN domain containing 1	0.47	23
11443	FBpp0079162	CG7429-PA	coiled-coil domain containing 53	0.47	32
12022	FBpp0084411	CG5484-PC, isoform C	Yip1 interacting factor homolog B	0.47	54
11551	FBpp0074964	CG6259-PA	charged multivesicular body protein 5	0.47	55
11887	FBpp0070637	CG6133-PA	NOL1/NOP2/Sun domain family, member 2	0.47	39
12058	FBpp0074616	FRG1 protein homolog		0.47	42
11533	FBpp0087458	CG12214-PA, isoform A	tubulin-specific chaperone e	0.47	29
12104	FBpp0085393	CG7791-PA	mitochondrial intermediate peptidase	0.47	34
11938	FBpp0083768	CG13827-PA	peroxisomal biogenesis factor 11 gamma	0.47	27
12004	FBpp0071818	Hypothetical UPF0172 protein CG3501.		0.47	38
11439	FBpp0085629	CG15087-PA	chromosome 11 open reading frame2	0.47	28
11361	FBpp0082332	CG3061-PA	DNA-J, putative	0.47	52
11779	FBpp0070924	COQ7		0.47	46
11421	FBpp0074715	anti-silencing factor 1		0.47	39

## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

11626	FBpp0076459	CG7550-PA	2-aminoethanethiol (cysteamine) dioxygenase	0.47	26
11650	FBpp0080553	Putative conserved oligomeric Golgi complex component 5		0.47	24
11715	FBpp0086992	CG18177-PB, isoform B		0.47	28
11981	FBpp0077333	CG3542-PB, isoform B	U1 small nuclear ribonucleoprotein putative	0.46	37
11961	FBpp0084418	CG6420-PA	WD-repeat protein	0.46	19
11517	FBpp0073082	CG14997-PB, isoform B	sulfide quinone reductase	0.46	52
11957	FBpp0089113	Transcription elongation factor SPT5		0.46	15
11414	FBpp0079258	CG12375-PA	metallo-beta-lactamase putative	0.46	47
11729	FBpp0079469	CG4537-PA	cysteine-rich PDZ-binding protein	0.46	29
11807	FBpp0081556	Spermidine Synthase		0.46	47
11464	FBpp0079697	CG6415-PA	aminomethyltransferase	0.46	39
12025	FBpp0082065	Aos1		0.46	41
11684	FBpp0083351	CG4159-PA	pseudouridylate synthase	0.46	43
12036	FBpp0072421	Enhancer of bithorax, isoform C		0.46	14
11743	FBpp0071597	CG9865-PB, isoform B	phosphatidylinositol glycan anchor biosynthesis, class M (CG9865)	0.46	32
11452	FBpp0083214	Vacuolar ATP synthase subunit G		0.46	102
11770	FBpp0085121	39S ribosomal protein L32, mitochondrial precursor		x 0.46	42
11856	FBpp0080638	CG12750-PA	cell cycle control protein cwf22	0.46	28
11433	FBpp0079468	FK506-binding protein 59		0.46	57
11623	FBpp0075393	CG6859-PA	peroxosomal biogenesis factor	0.46	36
11493	FBpp0085258	CG1416-PC, isoform C	AHA1, activator of heat shock 90kDa protein ATPase homolog 1	0.46	50
11522	FBpp0072564	CG9153-PB, isoform B	hect E3 ubiquitin ligase	0.46	36
11415	FBpp0072841	mitochondrial ribosomal protein S35		x 0.46	43
11374	FBpp0085363	SCAP		0.46	17
11365	FBpp0077150	Probable DNA replication complex GINS protein PSF2	GINS complex subunit 2 (Pst2 homolog)	0.46	33
11367	FBpp0070302	Myb-interacting protein 130		0.46	20
11494	FBpp0079203	CG8506-PA	zinc finger, FYVE domain containing 20	0.46	28
12001	FBpp0099494	C-1-tetrahydrofolate synthase, cytoplasmic		0.45	18
11977	FBpp0075120	CG4098-PA	nudix hydrolase 6	0.45	27
11766	FBpp0100136	1-phosphatidylinositol-4,5-bisphosphate phosphodiesterase		0.45	19
11947	FBpp0072946	CG11526-PB, isoform B	family with sequence similarity 40, member A	0.45	20
11877	AAEL006769-PA	tryptophanyl-tRNA synthetase		0.45	25
11670	FBpp0080918	CG2614-PA	KIAA0859	0.45	34
11946	FBpp0079699	CG6443-PA	chromosome 20 open reading frame 43	0.44	54
11453	FBpp0072961	CG14967-PA	KIAA0100	0.44	18
11679	FBpp0077525	tho2		0.44	24
11943	FBpp0078358	CG12170-PA	3-oxoacyl-[acyl-carrier-protein] synthase	0.44	36
11664	FBpp0074808	CG3808-PA	RNA m5u methyltransferase	0.44	40
11900	FBpp0073966	Clastrin heavy chain		0.44	18
11713	FBpp0073663	iodotyrosine dehalogenase	iodotyrosine dehalogenase	0.44	17
11562	FBpp0079897	CG6746-PA	ptpla domain protein	0.44	53
12086	FBpp0071031	Probable mitochondrial import receptor subunit TOM40 homolog		0.44	54
11748	FBpp0087244	CG30022-PA	beta lactamase domain	0.44	55
11655	FBpp0084069	tolkin, isoform B		0.44	16
11732	FBpp0084626	CG4849-PA	116 kDa U5 small nuclear ribonucleoprotein component	0.44	17
11738	FBpp0081355	CG9630-PA	DEAD box ATP-dependent RNA helicase	0.44	29
11505	FBpp0086380	CG8443-PA	eukaryotic translation initiation factor 3 subunit (eif-3)	0.43	16
11678	FBpp0072703	CG13926-PA	chromosome 11 open reading frame 73	0.43	37
11466	FBpp0084779	ligatin		0.43	26
11906	FBpp0081810	CG6608-PB, isoform B	mitochondrial carrier protein putative	0.43	32
11605	FBpp0076242	CG5026-PA, isoform A	myotubularin	0.43	28
11445	FBpp0085923	adipose		0.43	25
11948	FBpp0071194	Probable U3 small nucleolar RNA-associated protein 11		0.43	42
11901	FBpp0074131	Integrin alpha-PS2 precursor		0.43	17
11747	FBpp0073491	CG1824-PA	lipid a export ATP-binding/permease protein msba	0.43	26
11942	FBpp0099560	Protein retinal degeneration B		0.43	23
12119	FBpp0077133	CG17840-PA	inositol 5-phosphatase	0.42	26
11818	FBpp0084478	CG5880-PA	zinc finger, DHHC-type containing 16	0.42	26
11371	FBpp0077357	okra		0.42	23



## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

11845	FBpp0071543	CG30390-PA	coiled-coil domain containing 101		0.42	29
11397	FBpp0083272	Ire-1	Serine threonine-protein kinase endoplasmic reticulum to nucleus signaling 2		0.42	19
11979	FBpp0083132	gata			0.42	29
11510	FBpp0075990	1,2-dihydroxy-3-keto-5-methylthiopentene dioxxygenase	acireductone dioxxygenase		0.42	53
12064	FBpp0083137	CG14290-PB	brain protein 44-like		0.42	34
11472	FBpp0062817	CG16941-PA	spliceosome associated protein		0.42	19
11532	FBpp0070299	CG14805-PA	PAF acetylhydrolase 45 kDa subunit putative		0.42	42
11781	FBpp0073083	pavarotti			0.42	24
12124	FBpp0078448	Probable proteasome subunit beta type 4		ph, du, de, ba [psmb-N]	0.41	92
12100	AAEL010379-PA	ATP-binding cassette transporter			0.41	16
12089	FBpp0072366	3-phosphoinositide-dependent protein kinase 1			0.41	34
12020	FBpp0075609	CG11267-PA	heat shock protein putative	du [Heat shock 10 kDa protein 1 (chaperonin 10)]	0.41	83
11599	FBpp0076280	CG5288-PC, isoform C	galactokinase		0.41	41
11462	FBpp0087323	CG6751-PA	WD-repeat protein		0.41	38
11492	FBpp0110411	conserved hypothetical protein			0.41	18
11949	FBpp0086399	CG8397-PA	actin binding protein putative		0.41	60
11662	FBpp0075280	Homeotic gene regulator			0.41	22
11485	FBpp0083354	Elongin B		du [elongin B isoform a]	0.41	57
11582	AAEL004330-PA	conserved hypothetical protein			0.41	17
11369	FBpp0085431	Transcription-associated protein 1 (Nipped-A)	transformation/transcription domain-associated protein		0.41	15
11986	FBpp0071155	Neuroglian precursor			0.41	15
11969	FBpp0070319	CG4199-PA, isoform A	disulfide oxidoreductase		0.41	34
11486	FBpp0110523	nitrate, fromate, iron dehydrogenase			0.41	34
11765	AAEL011712-PA	diacylglycerol kinase			0.41	12
11973	FBpp0075344	CG7650-PA	viral IAP-associated factor putative		0.41	41
11994	FBpp0072767	CG8993-PA	thioredoxin putative		0.41	61
11758	FBpp0076708	Transportin, isoform A			0.41	19
11682	FBpp0085071	Protein tailless			0.41	16
11638	FBpp0073725	CG1461-PA	tyrosine aminotransferase		0.41	40
11661	FBpp0070304	CG3573-PA	inositol polyphosphate 5-phosphatase		0.40	27
11746	FBpp0088517	CG5009-PA	acyl-CoA oxidase		0.40	28
11933	FBpp0084307	CG4743-PA	mitochondrial carrier protein		0.40	26
11475	FBpp0073983	Actin-like protein 13E			0.40	32
12052	FBpp0080894	cdc23			0.40	27
11689	FBpp0087297	BBS4			0.40	20
11717	FBpp0075685	Protein angel			0.40	24
12030	FBpp0072119	CG3735-PA	chromosome 1 open reading frame 107		0.40	29
11999	FBpp0079776	CG6700-PA	leukocyte receptor cluster (lrc) member		0.40	29
11888	FBpp0085589	CG11788-PA	defective in sister chromatid cohesion 1 homolog		0.40	34
11780	FBpp0080922	Importin beta subunit			0.40	19
11568	FBpp0074481	CG12203-PA	NADH:ubiquinone dehydrogenase putative	du [NADH dehydrogenase (ubiquinone) Fe-S protein 4]	0.40	72
11897	FBpp0082657	Mitochondrial import inner membrane translocase subunit TIM16			0.40	43
11597	FBpp0087591	Protein preli-like			0.40	55
11540	FBpp0087891	CG8709-PA	lipin		0.40	32
11648	FBpp0080807	Probable phosphomevalonate kinase			0.39	37
11950	FBpp0088810	Protein arginine N-methyltransferase capsuleen			0.39	37
11554	FBpp0071033	Pyruvate dehydrogenase phosphatase (CG12151-PA)			0.39	29
11468	FBpp0079547	Niemann-Pick Type C-1			0.39	24
11611	AAEL006321-PA	1-acylglycerol-3-phosphate acyltransferase			0.39	20
11967	FBpp0078711	CG12512-PA	AMP dependent coa ligase		0.39	43
12072	FBpp0082148	CG5608-PA	Vac14 homolog		0.39	33
11541	FBpp0079586	CG31715-PA	myotrophin		0.39	37
11968	FBpp0087979	Cytochrome b5			0.38	89
11395	FBpp0076789	Pole2			0.38	27
12117	FBpp0081376	Dihydroorotate dehydrogenase, mitochondrial precursor			0.38	34
11927	FBpp0081157	CG1104-PA, isoform A			0.38	27
11359	FBpp0083514	DNA polymerase alpha catalytic subunit			0.38	17
11966	FBpp0082813	CG3534-PA	xylulose kinase		0.38	30
11876	FBpp0074672	Translocase of outer membrane 20			0.38	67

## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

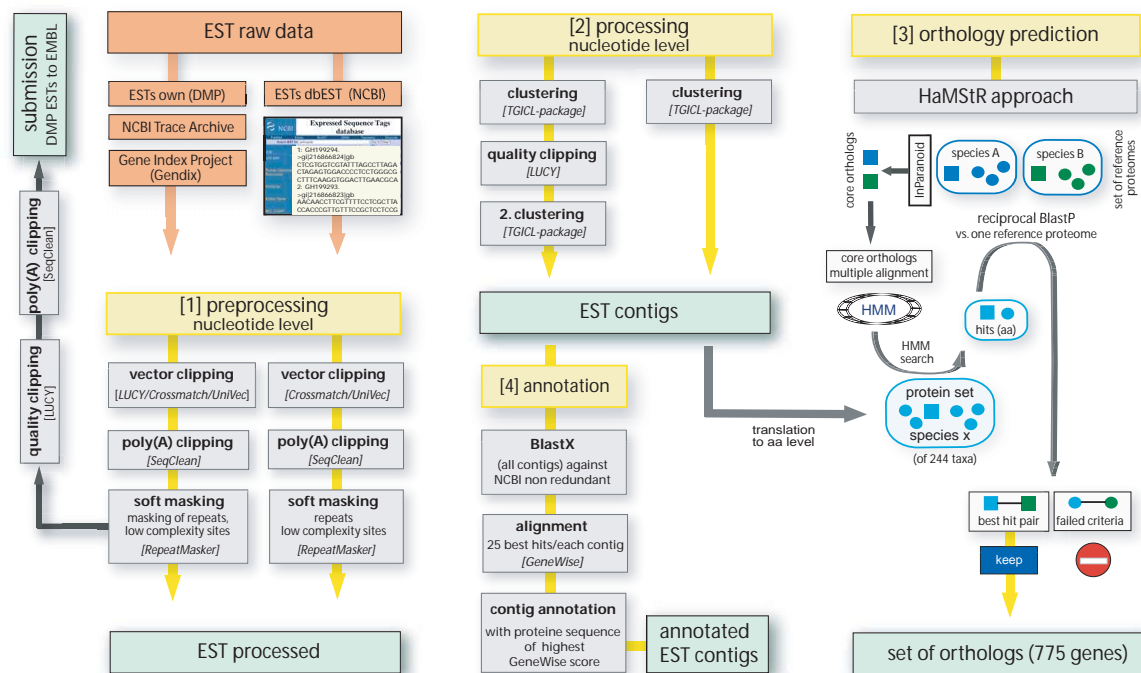
11908	FBpp0079488	CG13126-PA	<i>methyltransferase 11 domain containing 1</i>		0.38	30
11497	FBpp0110309	poly a polymerase			0.38	21
11401	FBpp0083840	CG10365-PA, isoform A	<i>ChaC, cation transport regulator homolog 1</i>		0.38	38
11842	FBpp0073355	Probable signal peptidase complex subunit 2		du [signal peptidase complex subunit 2 homolog]	0.38	67
11791	FBpp0077447	CG9867-PA	glycosyltransferase		0.38	29
11872	FBpp0078684	CG8891-PA	inosine triphosphate pyrophosphatase (itpase) (inosine triphosphatase)		0.37	37
11470	FBpp0084728	Protein kinase C			0.37	21
12095	FBpp0070301	mitochondrial ribosomal protein L16		x	0.37	41
12041	FBpp0074227	CG5800-PA	DEAD box ATP-dependent RNA helicase		0.37	34
11923	FBpp0083244	CG4973-PA	zinc finger protein putative		0.37	35
11885	FBpp0084711	CG1951-PA	SCY1-like 2		0.37	21
11622	FBpp0083022	CG7146-PA	vacuolar protein sorting 39 homolog		0.37	26
11498	FBpp0081588	CG9399-PA, isoform A	<i>brain protein 44</i>		0.37	60
11353	FBpp0074004	CG32579-PA	<i>XK, Kell blood group complex subunit-related family, member 6</i>		0.37	21
11870	FBpp0079567	CG31717-PA	<i>phosphatidic acid phosphatase type 2 domain containing 2</i>		0.37	39
11951	FBpp0077965	UPF0315 protein			0.37	43
11560	FBpp0078275	jagunal, isoform C			0.37	47
11690	FBpp0077209	Pdsw, isoform B			0.37	71
11777	FBpp0087085	DNA-directed RNA polymerase III 128 kDa polypeptide			0.36	16
12091	FBpp0083854	Probable oligoribonuclease			0.36	36
11593	FBpp0081841	CG17187-PA	<i>DnaJ (Hsp40) homolog, subfamily C, member 17</i>		0.36	33
11851	FBpp0071891	Arginine methyltransferase 7			0.36	34
11581	FBpp0073235	Putative 6-phosphogluconolactonase	6-phosphogluconolactonase		0.36	47
11481	FBpp0086877	CG4646-PA	<i>chromosome 1 open reading frame 123</i>		0.36	36
12032	FBpp0083853	twister			0.35	17
11569	FBpp0081451	Adenosine deaminase			0.35	29
11703	FBpp0080628	CG15161-PA			0.35	21
11476	FBpp0071426	CG1826-PA	<i>BTB (POZ) domain containing 9</i>		0.35	29
11749	FBpp0078844	CG9154-PA	<i>N-6 adenine-specific DNA methyltransferase 2 (putative)</i>		0.35	34
11706	FBpp0082314	CG9588-PA	26S proteasome non-ATPase regulatory subunit		0.35	48
11612	FBpp0073995	CG3560-PA	ubiquinol-cytochrome c reductase complex 14 kd protein		0.35	85
12034	FBpp0076111	Laminin gamma-1 chain precursor			0.35	18
11548	FBpp0074104	mitochondrial ribosomal protein L22		x	0.35	44
11620	FBpp0073585	Vesicular-fusion protein Nsf1			0.34	20
11557	FBpp0079620	CG6206-PB, isoform B	lysosomal alpha-mannosidase (mannosidase alpha class 2b member 1)		0.34	42
11543	FBpp0089008	Adenine phosphoribosyltransferase			0.34	49
11602	AAEL007823-PA	PIWI			0.34	18
12015	FBpp0085924	CG10914-PA			0.34	26
11931	FBpp0089163	Cleavage and polyadenylation specificity factor, 160 kDa subunit			0.34	19
11696	FBpp0082172	Xanthine dehydrogenase			0.34	31
11399	FBpp0086640	DNA-directed RNA polymerase I largest subunit	DNA-directed RNA polymerase I largest subunit		0.33	20
12006	FBpp0080203	DNA mismatch repair protein spellchecker 1			0.33	22
11963	FBpp0086591	SMC2			0.33	23
11744	FBpp0073979	Graf, isoform A			0.33	25
11440	FBpp0078583	CG9804-PA	lipote-protein ligase b		0.33	29
11575	FBpp0084349	Dak1			0.33	56
11594	FBpp0086887	Tripeptidyl-peptidase 2			0.33	19
11832	FBpp0072460	Rhythmically expressed gene 2 protein			0.32	25
12113	FBpp0075106	Probable ATP-dependent RNA helicase Dbp73D			0.32	32
11585	FBpp0071193	Hypothetical protein CG1785			0.32	41
11693	FBpp0083857	Putative succinate dehydrogenase [ubiquinone] cytochrome b small subunit, mitochondrial precursor			0.32	63
11985	FBpp0076589	Signal recognition particle 19 kDa protein	srp19		0.32	50
11751	FBpp0081763	CG4511-PA	viral IAP-associated factor putative		0.32	51
11595	FBpp0075755	lethal (3) neo18		du [NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 5, 1 kDa subunit]	0.32	73
11633	FBpp0075399	Probable DNA mismatch repair protein MSH6			0.32	25
12021	FBpp0072426	thoc7, isoform A			0.32	41
11410	FBpp0070403	Probable ATP-dependent RNA helicase	ATP-dependent RNA helicase		0.31	20
12005	FBpp0080475	CG31739-PA	aspartyl-tRNA synthetase		0.31	27
11658	FBpp0081860	mitochondrial ribosomal protein L40		x	0.31	46

## Appendix 6: A phylogenomic approach to resolve the arthropod tree of life

11420	FBpp0082522	ATP synthase O subunit, mitochondrial precursor		du [Mitochondrial ATP synthase, O subunit precursor]	0.31	111
11503	FBpp0075148	CG33158-PB	translation elongation factor elongation factor Tu GTP binding domain containing 1 isoform 2		0.31	28
11819	FBpp0077251	CG33123-PA	leucyl-tRNA synthetase		0.30	18
11504	FBpp0085690	CG11242-PA	tubulin-specific chaperone b (tubulin folding cofactor b)		0.30	48
11373	FBpp0078895	CG9542-PA	arylfornamidase		0.30	28
11430	FBpp0086226	Superoxide dismutase [Mn], mitochondrial precursor			0.30	81
11879	FBpp0070639	CG6379-PA	<i>FtsJ</i> methyltransferase domain containing 2		0.30	26
11702	FBpp0071916	CG11079-PC, isoform C	5-formyltetrahydrofolate cyclo-ligase		0.30	41
11537	FBpp0083226	CG4686-PA			0.30	51
11945	FBpp0073762	Probable mitochondrial 28S ribosomal protein S25		x	0.29	45
12044	FBpp0073196	CG15014-PA	<i>THUMP</i> domain containing 1		0.29	42
11800	FBpp0077399	Transportin-Serine/Arginine rich			0.29	23
12028	FBpp0077173	CG31961-PA, isoform A	tubulin folding cofactor c		0.28	36
11873	AAEL011682-PA	nuclear pore complex protein nup93			0.28	17
11907	FBpp0083650	Probable prefoldin subunit 5			0.27	63
11519	FBpp0072615	CG9187-PA	partner of slid5		0.27	32
11388	FBpp0087629	CG1884-PB, isoform B			0.27	16
11736	FBpp0084051	CG13625-PA	<i>BUD13</i> homolog		0.27	34
11812	FBpp0100031	Protein male-less	ATP-dependent RNA helicase		0.26	26
11753	FBpp0079316	CG13397-PA	alpha-n-acetylglucosaminidase		0.26	22
12055	FBpp0072456	Rev1			0.25	20
12027	AAEL011963-PA	conserved hypothetical protein			0.25	18
11836	AAEL009888-PA	WD-repeat protein			0.25	25
11665	AAEL004081-PA	dj-1 protein			0.25	57
11995	FBpp0080305	CG15261-PA	ribonuclease UK114 putative		0.24	70
11516	AAEL005494-PA	conserved hypothetical protein			0.17	9

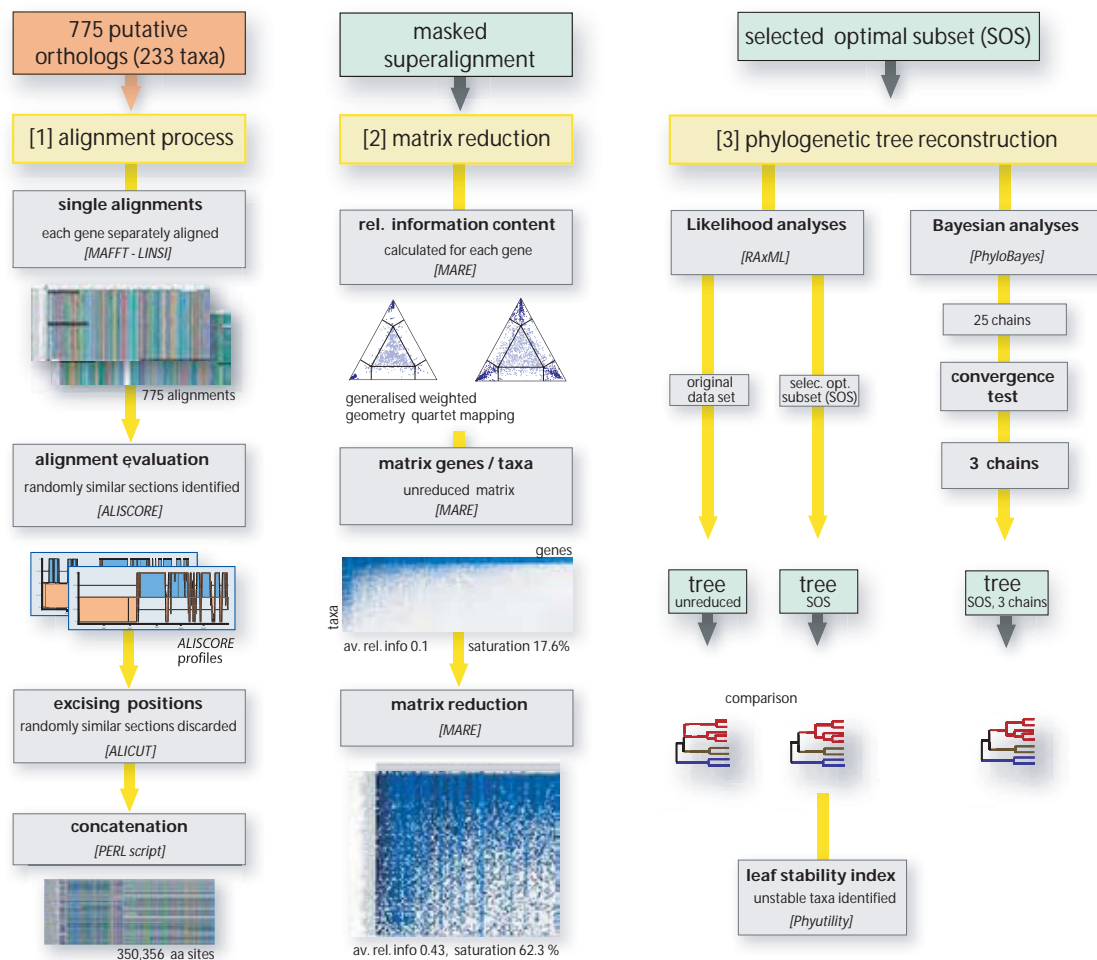
### Supplementary Figure 1 | Processing of EST data and orthology assignment.

EST raw data (orange) of own and public EST projects were mined and processed in four major steps (yellow), preprocessing, processing, orthology prediction and annotation. [1] In the preprocessing own EST sequences were screened for vectors and poly(A) tails using *LUCY*<sup>9</sup>. All sequences including published ESTs were screened for contamination by comparison against the data base *UniVec* (<http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html>) with *Crossmatch*<sup>10</sup> ([http://www.incogen.com/public\\_documents/vibe/details/crossmatch.html](http://www.incogen.com/public_documents/vibe/details/crossmatch.html)) and *SeqClean*<sup>11</sup> (<http://www.tigr.org/tdb/tgi/software/>). *SeqClean* screened for poly(A) tails as well. We discarded sequences < 100 nucleotides. Afterwards repetitive elements in the remaining ESTs were soft masked with RepeatMasker<sup>12</sup> using *Repbse*<sup>13</sup>. [2] ESTs were clustered using the *TGICL* package. ESTs of own projects were quality clipped with *LUCY* and clustered a second time to obtain and keep longer sequences for the EST contigs. The contigs were translated into amino acid level and [3] integrated in the orthology prediction with HaMStR<sup>1</sup>. We compiled a set of reference proteomes (InParanoid<sup>2-4</sup>, <http://inparanoid6.sb.su.se>) with *D. pulex*, *T. castaneum*, *B. mori*, *A. aegypti*, *A. mellifera*, *D. melanogaster*, *C. elegans*, *C. briggsae*, *Capitella sp.*, *L. gigantea*, *H. sapiens*, *T. nigroviridis* and *X. tropicalis* as 'primer' taxa. Multiple alignments of 'core' orthologs for primer taxa were used to train Hidden Markov Models (HMMs) to search in each protein set of our 244 taxa for hits. A reciprocal BlastP<sup>14</sup> decides about the surviving of a hit. For the re-blast step we always chose the proteome of the presumably evolutionary closest primer taxon for each considered species. We upended in the set of 775 putative orthologous gene loci. [4] EST contigs were annotated using a BlastX search against NCBI's non-redundant protein database. The protein sequences of the 25 best hits for each contig were aligned with *GeneWise*<sup>15</sup>. The contig is annotated according to the protein sequence with the highest *GeneWise* score. Single EST reads (supplementary table 1) were submitted to EMBL.



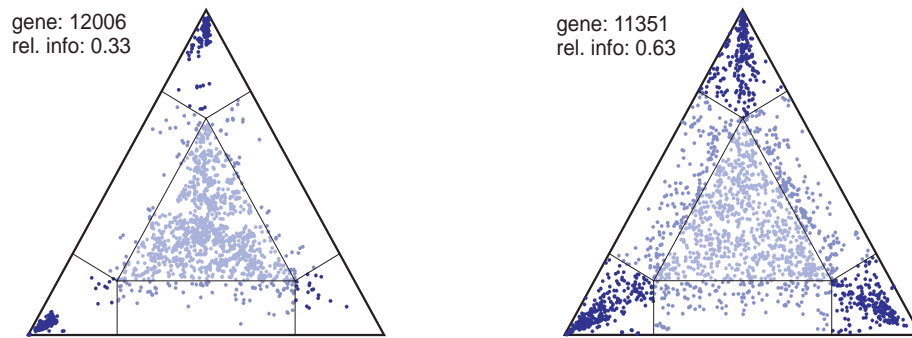
## Supplementary Figure 2 | Alignment masking, selecting an optimal data subset and phylogenetic analyses.

Based on the 775 putative orthologs (orange) the working flow consists of three major steps (yellow): alignment process, reduction heuristics and phylogenetic reconstruction. [1] The alignment process starts aligning each single gene separately using MAFFT<sup>16</sup>. ALISCORE (<http://aliscore.zfmk.de>) identifies randomly similar sections in each alignment, ALICUT (<http://utilities.zfmk.de>) discards by ALISCORE negatively scored positions. The genes are concatenated to a masked superalignment (green). [2] The step of reduction heuristics starts with the calculation of the relative information content of each gene in the masked superalignment. The generated matrix taxa vs. genes is given with a value for relative information content of each gene. An optimal subset is selected (SOS) by excluding genes and taxa showing low relative information content (see methods). [3] Phylogenetic trees were constructed using RAxML<sup>17,18,19</sup> and PhyloBayes<sup>20,21</sup>. The two resulting ML trees (green) base on the original data set and the selected optimal subset (SOS). Phyutility<sup>22</sup> was used to identify ‘unstable’ taxa. In Bayesian analyses we ran 25 chains. After testing for topological incongruences (see methods) we inferred a ‘triple’ majority rule consensus tree from 3 chains.



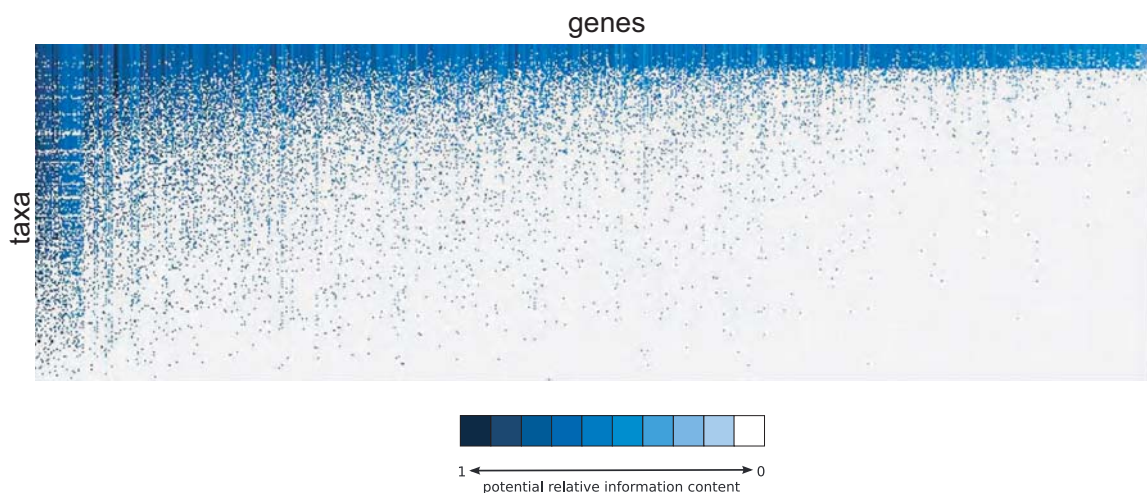
**Supplementary Figure 3 | Potential relative information content of genes visualised by 2D simplex bipartite graphs.**

Potential information content (rel. info) of a single partition (gene) is defined as the relative treelikeness of the data using geometry mapping<sup>23</sup>, extended to amino acids incorporating the BLOSUM62 substitution matrix. Relative tree-likeness corresponds to the relative frequency of simplex points within the outer areas of at least partially resolved trees compared to the total number of simplex points. Genes containing less than four sequences and taxa containing less than 1/3 of the single gene sequence are considered as absent.



**Supplementary Figure 4 | Original data matrix with potential relative information content of each gene and taxon.**

The matrix comprises 233 taxa (rows) and 775 genes (columns). Potential relative information content ranges from 0.0 – 1.0 (10 units). Potential relative information content is color coded from dark blue (> 0.9 – 1.0) to white (relative information content of 0 - 0.1 or missing data). Genes with a relative information content < 0.04 were considered as absent. Overall average relative information content of the matrix: 0.1, overall saturation: 17.6%.









## Supplementary Material: References

1. Ebersberger, I., Strauss, S. & von Haeseler, A. HaMStR Profile Hidden Markov Model based search for orthologs in ESTs. *BMC Evol. Biol.* **9**, 157, doi: 10.1093/molbev/msp191(2009).
2. Remm, M., Storm, C. E. V. & Sonnhammer, E. L. L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**, 1041–1052 (2001).
3. Berglund, A.-C., Sjölund, E., Östlund, G. & Sonnhammer, E. L. L. InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res.* **36** (Database issue), D263–266 (2008).
4. Berglund, A. C., Sjölund, E., Östlund, G. & Sonnhammer, E. L. L. InParanoid version6.1. <http://inparanoid6.sbc.su.se/cgi-bin/index.cgi> (2008).
5. Philippe, H. *et al.* Phylogenomics revives traditional views on deep animal relationships. *Curr. Biol.* **19**, 706–712 (2009).
6. Delsuc, F., Tsagkogeorga, G., Lartillot, N. & Philippe, H. Additional molecular support for the new chordate phylogeny. *Genesis* **46**, 592–604 (2008).
7. Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745–749 (2008).
8. Baurain, D., Brinkmann, H. & Philippe, H. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? *Mol. Biol. Evol.* **24**, 6– 9 (2007).
9. Chou, H. H. & Holmes, M. H. DNA sequence quality trimming and vector removal. *Bioinformatics* **17**, 1093–1104 (2001).
10. Green, P. Crossmatch. [http://www.incogen.com/public\\_documents/vibe/details/crossmatch.html](http://www.incogen.com/public_documents/vibe/details/crossmatch.html) (1993-1996).
11. SeqClean. <http://www.tigr.org/tdb/tgi/software/> (2005-2006).
12. Smit, A.F.A., Hubley, R. & Green, P. RepeatMasker Open-3.0. <http://www.repeatmasker.org> (1996-2004).
13. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
14. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of proteindatabase search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
15. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
16. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinformatics* **9**, 286–298 (2008).
17. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
18. Ott, M., Zola, J., Aluru, S. & Stamatakis, A. Large-scale maximum likelihood-based phylogenetic analysis on the IBM BlueGene/L. In *Proceedings of the 2007 ACM/IEEE conference on Supercomputing*. Reno, Nevada, 2007 (ACM, New York, NY, USA).
19. Stamatakis, A. Phylogenetic models of rate heterogeneity: a high performance computing perspective. In *Proceedings of the Parallel and Distributed Processing Symposium (IPDPS) 2006*. Rhodes, Greece (2006).
20. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109 (2004).
21. Lartillot, N., Blanquart, S., & Lepage, T. PhyloBayes 2.3 - a Bayesian software for phylogenetic reconstruction using mixture models. <http://www.lirmm.fr/mab/IMG/pdf/phylobayes2.3.pdf> (2007).
22. Smith, S. A. & Dunn, C. W. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24**, 715–716 (2008).
23. Nieselt-Struwe, K. & von Haeseler, A. Quartet-mapping, a generalization of the likelihood-mapping procedure. *Mol. Biol. Evol.* **18**, 1204–1219 (2001).

**Additional file 1:** Primer list for *Baetis* sp.

Primer	Direction	Gene	Sequence (5'-3')	Source	Notes
16Sfw	forward	16S rRNA	GCTAGAATCCTAGGTATTGCCTGCC	S.Simon	<i>Baetis</i> sp. specific
16Srev	reverse	16S rRNA	TCTACGGGGTCTTCTCGTCCTGC	S.Simon	<i>Baetis</i> sp. specific
Cytfw	forward	<i>cytb</i>	GTTGGCTGCTCCGAACGTTACATGC	S.Simon	<i>Baetis</i> sp. specific
Cytrrev	reverse	<i>cytb</i>	ATCAACTGCAAAGCCTCCTCAAACC	S.Simon	<i>Baetis</i> sp. specific
NADH4fw	forward	<i>nad4</i>	CACGTAGAGGCACCTGTAAGAGG	S.Simon	<i>Baetis</i> sp. specific
NADH4rev	reverse	<i>nad4</i>	CTTATGTGAGCTACTGAGGAGTAAGC	S.Simon	<i>Baetis</i> sp. specific
NADH5fw	forward	<i>nad5</i>	GGTTGGGATGGTTTGGGTTTAGTATCC	S.Simon	<i>Baetis</i> sp. specific
NADH5rev	reverse	<i>nad5</i>	GTAACCAAGCTGAAAATGGGATTTGAGC	S.Simon	<i>Baetis</i> sp. specific
COX2fw	forward	<i>cox2</i>	TTATCGCCTTACCATCTCTGCGG	S.Simon	<i>Baetis</i> sp. specific
COX2rev	reverse	<i>cox2</i>	GGCACACCGTCGACCTTTACACC	S.Simon	<i>Baetis</i> sp. specific
COX1rev	reverse	<i>cox1</i>	GGATCCCCACCTCCAGCAGG	S.Simon	<i>Baetis</i> sp. specific

**Additional file 2:** PCR conditions for *Baetis* sp.

Fragment 16SrRNA - *cytb*

Bioline, Taq Polymerase in $\mu\text{L}$	
Water	14,75
10x Buffer (20mM)	2,5
MgCl <sub>2</sub> (50mM)	1,5
dNTPS (2,5mM)	2,5
Primer forward (10 $\mu\text{M}$ )	1,25
Primer reverse (10 $\mu\text{M}$ )	1,25
Template	1
Taq Polymerase (5u/ $\mu\text{L}$ )	0,25
Total	25

Thermocycler: 9700 (Applied Biosystems)

Program for amplification		
Initial denaturation	95°C	2,5 min
Denaturation	94°C	30 sec
Annealing	59,3°C	30 sec
Elongation	72°C	3,5 min
Final Elongation	72°C	2 min
Number of cycles 45		

Fragment *cytb* - *nad4*

Bioline, Taq Polymerase in $\mu\text{L}$	
Water	14,75
10x Buffer (20mM)	2,5
MgCl <sub>2</sub> (50mM)	1,5
dNTPS (2,5mM)	2,5
Primer forward (10 $\mu\text{M}$ )	1,25
Primer reverse (10 $\mu\text{M}$ )	1,25
Template	1
Taq Polymerase (5u/ $\mu\text{L}$ )	0,25
Total	25

Thermocycler: 9700 (Applied Biosystems)

Program for amplification		
Initial denaturation	95°C	2,5 min
Denaturation	94°C	30 sec
Annealing	55,9°C	30 sec
Elongation	72°C	3,5 min
Final Elongation	72°C	2 min
Number of cycles 45		

Fragment *nad4* - *nad5*

Bioline, Taq Polymerase in $\mu\text{L}$	
Water	14,75
10x Buffer (20mM)	2,5
MgCl <sub>2</sub> (50mM)	1,5
dNTPS (2,5mM)	2,5
Primer forward (10 $\mu\text{M}$ )	1,25
Primer reverse (10 $\mu\text{M}$ )	1,25
Template	1
Taq Polymerase (5u/ $\mu\text{L}$ )	0,25
Total	25

Thermocycler: 9700 (Applied Biosystems)

Program for amplification		
Initial denaturation	95°C	2,5 min
Denaturation	94°C	30 sec
Annealing	61,0°C	30 sec
Elongation	72°C	3,5 min
Final Elongation	72°C	2 min
Number of cycles 45		

Fragment *nad5-cox2*

Roche, Expand Longe Range in $\mu\text{L}$	
Water	27,8
DMSO	2
dNTPS	2,5
5x Buffer (+ MgCl <sub>2</sub> )	10
Primer forward (10 $\mu\text{M}$ )	2,5
Primer reverse (10 $\mu\text{M}$ )	2,5
Template	2
Taq Polymerase	0,7
Total	50

Thermocycler: 9700 (Applied Biosystems)

Program for amplification		
Initial denaturation	94°C	2 min
Denaturation	94°C	10 sec
Annealing	57°C	15 sec
Elongation	68°C	10 min
Denaturation	94°C	10 sec
Annealing	57°C	5 sec
Elongation	68°C	10 min +20
Final Elongation	68°C	7 min
Step 1 = 10 cycles step 2 = 35 cycles		

Fragment *cox2-cox1*

Bioline, Taq Polymerase in $\mu\text{L}$	
Water	14,85
10x Buffer (20mM)	2,5
MgCl <sub>2</sub> (50mM)	1,5
dNTPS (2,5mM)	2,5
Primer forward (10 $\mu\text{M}$ )	1,25
Primer reverse (10 $\mu\text{M}$ )	1,25
Template	1
Taq Polymerase (5u/ $\mu\text{L}$ )	0,15
Total	25

Program for amplification		
Initial denaturation	95°C	2,5 min
Denaturation	94°C	30 sec
Annealing	60°C	30 sec
Elongation	72°C	2,0 min
Final elongation	72°C	2,0 min
Number of cycles 45		

**Additional file 3:** Primer list for *Boyeria irene*..

Primer	Direction	Gene	Sequence (5'-3')	Source	Notes
TK-N-3785	reverse	cox2	GTTTAAGAGACCAGTACTTG	Simon et al., 1994	universal
C2-J-3400	forward	cox2	ATTGGACATCAATGATATTGA	Simon et al., 1994	universal
P2216	reverse	16S rRNA	TAATCCAACATCGAGGTCGCAA	A.Wargel	universal
P2215	forward	16S rRNA	GACCGTGCRAGGATAGCATAATCA	A.Wargel	universal
P2246	forward	16S rRNA	TGGAAGACGAGAAGACCCTATAGAGC	S.Dellaporta	<i>B.irene</i> specific
P2246comp	reverse	16S rRNA	GCTCTATAGGGTCTTCTCGTCTTCCA	S.Dellaporta	<i>B.irene</i> specific
P3141	forward	cox2	AAGTAGATGCCACTCCTGGTCGATT	S.Dellaporta	<i>B.irene</i> specific
P3141comp	reverse	cox2	AATCGACCAGGAGTGGCATCTACTT	S.Dellaporta	<i>B.irene</i> specific
cox3revb.i.	reverse	cox3	CATCAACAAAGTGTCAATATCACGC	S.Simon	<i>B.irene</i> specific
nad3fwb.i.	forward	nad3	CTCCATTTGAATGTGGATTTGATCC	S.Simon	<i>B.irene</i> specific
nd5fwb.i.	forward	nad5	TATTAGGTTGGGATGGATTGG	S.Simon	<i>B.irene</i> specific
nad5revb.i2	reverse	nad5	CCAATCCATCCCAACCTAATA	S.Simon	<i>B.irene</i> specific
cytbfbw.i.	forward	cytb	CCTGCAAATCCTTTAGTAACGCC	S.Simon	<i>B.irene</i> specific
cytbrevb.i.2	reverse	cytb	GGCGTTACTAAAGGATTTGCAGG	S.Simon	<i>B.irene</i> specific
221-1-1F	forward	nad5-cytb	GAGTATAGGCAGCACTAAAAAATG	Macrogen	Sequencing primer
221-1-1R	reverse	nad5-cytb	CCCATAGTTTACATCACGACAAATG	Macrogen	Sequencing primer
221-1-2F	forward	nad5-cytb	CAACATGAGCCTTCGGTAATC	Macrogen	Sequencing primer
221-1-2R	reverse	nad5-cytb	CAAGTAGTATATCCTATGCAACTTG	Macrogen	Sequencing primer

Simon C, Frai F, Bechenback A, Crespi B, Liu H, Flook PK. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequence and a compilation of conserved polymerase chain reaction primers. Ann Entomol Soc Am 1994, 87: 651-701

**Additional file 4:** PCR conditions for *Boyeria irene*.

Fragment 16S rRNA - cox2 (~10kb)

TaKara LA HS in µL	
Water	15,25
10x Buffer (25mM)	2,5
dNTPS (2.5mM)	4
Primer forward (10µM)	1
Primer reverse (10µM)	1
Template	1
Taq Polymerase (5u/µL)	0,25
Total	25

Thermocycler: 9700 (Applied Biosystems)

Program for amplification		
Initial denaturation	94°C	2 min
Denaturation	94°C	30 sec
Elongation	67°C	10 min
Denaturation	94°C	30 sec
Annealing/Elongation	65°C	10 min
Final Elongation	68°C	7 min
Step 1 = 7 cycles step 2 = 30 cycles		

Fragment 16S rRNA - cox2 (~6kb)

NEB LongAmp Taq in µL	
Water	15
5x Buffer	5
dNTPS (10mM)	1
Primer forward (10µM)	1
Primer reverse (10µM)	1
Template	1
Taq Polymerase (5u/50µL)	1
Total	25

Thermocycler: 9700 (Applied Biosystems)

Program for amplification		
Initial denaturation	95°C	4 min
Denaturation	95°C	1 min
Elongation	60°C	10 min
Final Elongation	65°C	10 min
Number of cycles 40		

# Additional file 5

Taxa list of mitochondrial genomes of hexapods with GenBank accession numbers. \* indicates taxa included in reduced data set.

group	order	Species	GenBank accession no	reduced data set
Primary wingless hexapods	Collembola	<i>Cryptopygus antarcticus</i>	NC_010533	
		<i>Friezea grisea</i>	NC_010535	
		<i>Onychiurus orientalis</i>	NC_006074	
		<i>Orchesella villosa</i>	NC_010534	
		<i>Podura aquatica</i>	NC_006075	
		<i>Sminthurus viridis</i>	NC_010536	
		<i>Tetrodontophora bielanensis</i>	NC_002735	
		<i>Campodea fragilis</i>	NC_008233	
		<i>Campodea lubbocki</i>	NC_008234	
		<i>Japyx solifugus</i>	NC_007214	
	Archaeognatha	<i>Nesomachilis australica</i>	NC_006895	*
		<i>Petrobius brevistylis</i>	NC_007688	
		<i>Trigoniophthalmus alternatus</i>	NC_010532	*
	Zygentoma	<i>Pedetontus silvestrii</i>	NC_011717	
		<i>Thermobia domestica</i>	NC_006080	*
		<i>Atelura formicaria</i>	NC_011197	*
		<i>Tricholepidion gertschi</i>	NC_005437	*
"Palaeoptera"	Odonata	<i>Pseudolestes mirabilis</i>	FJ606784	*
		<i>Boyeria irene</i>	this study	*
		<i>Orthetrum triangulare</i>	AB126005	*
		<i>Davidius lunatus</i>	NC_012644	*
	Ephemeroptera	<i>Baetis</i> sp.	this study	*
		<i>Parafronurus youi</i>	NC_011359	*
		<i>Siphonurus immanis</i>	NC_013822	*
		<i>Ephemeria orientalis</i>	NC_012645	*
	Polyneoptera	<i>Periplaneta fuliginosa</i>	NC_006076	*
		<i>Blattella germanica</i>	NC_012901	*
Polyneoptera	Blattodea	<i>Locusta migratoria</i>	NC_001712	
		<i>Oedaleus decorus</i>	NC_011115	
	Caelifera	<i>Gastrimargus marmoratus</i>	NC_011114	
		<i>Oxya chinensis</i>	NC_010219	*
		<i>Prumna arctica</i>	NC_013835	
		<i>Arcyptera coreana</i>	NC_013805	
		<i>Schistocerca gregaria</i>	NC_013240	
		<i>Phlaeoba albonema</i>	NC_011827	*
		<i>Atractomorpha sinensis</i>	NC_011824	
		<i>Calliptamus italicus</i>	NC_011305	
		<i>Acrida willemsei</i>	NC_011303	
		<i>Gryllotalpa orientalis</i>	NC_006678	
	Ensifera	<i>Ruspolia dubia</i>	NC_009876	*
		<i>Teleogryllus emma</i>	NC_011823	
		<i>Deracantha onos</i>	NC_011813	*
		<i>Troglophilus neglectus</i>	NC_011306	
		<i>Gryllotalpa pluvialis</i>	NC_011302	
		<i>Myrmecophilus manni</i>	NC_011301	
		<i>Anabrus simplex</i>	NC_009967	
		<i>Reticulitermes flavipes</i>	NC_009498	
		<i>Reticulitermes hageni</i>	NC_009501	
		<i>Reticulitermes santonensis</i>	NC_009499	*
	Mantodea	<i>Reticulitermes virginicus</i>	NC_009500	
		<i>Tamolanica tamolana</i>	NC_007702	*
	Mantophasmatodea	<i>Sclerophasma paresisense</i>	NC_007701	*
	Phasmatodea	<i>Ramulus hainanense</i>	NC_013185	*
	Plecoptera	<i>Pteronarcys princeps</i>	NC_006133	*
Paraneoptera	Auchenorrhyncha	<i>Philaenus spumarius</i>	NC_005944	
		<i>Lycorma delicatula</i>	NC_012835	
		<i>Geisha distinctissima</i>	NC_012617	
		<i>Laodelphax striatellus</i>	NC_013706	
		<i>Valentia hoffmanni</i>	NC_012823	
		<i>Paraplea frontalis</i>	NC_012822	
	Heteropterida	<i>Ochterus marginatus</i>	NC_012820	
		<i>Enithares tibialis</i>	NC_012819	
		<i>Laccotrephes robustus</i>	NC_012817	
		<i>Yemmalysus parallelus</i>	NC_012464	
		<i>Saldula arsenjevi</i>	NC_012463	
		<i>Riptortus pedestris</i>	NC_012462	
		<i>Triatoma dimidiata</i>	NC_002609	

group	order	Species	GenBank accession no	reduced data set
Holometabola	Heteroptera	<i>Phaenacantha marcida</i>	NC_012460	
		<i>Neuroctenus parus</i>	NC_012459	
		<i>Malcus inconspicuus</i>	NC_012458	
		<i>Macroscytus subaeneus</i>	NC_012457	*
		<i>Hydaropsis longirostris</i>	NC_012456	
		<i>Coptosoma bifaria</i>	NC_012449	
		<i>Aeschyntelus notatus</i>	NC_012446	
		<i>Physopelta gutta</i>	NC_012432	
		<i>Orius niger</i>	NC_012429	
		<i>Geocoris pallidipennis</i>	NC_012424	
		<i>Dysdercus cingulatus</i>	NC_012421	
		<i>Nezara viridula</i>	NC_011755	
		<i>Halyomorpha halys</i>	NC_013272	
		<i>Stictopleurus subviridis</i>	NC_012888	
		<i>Ilyocoris cimicoides</i>	NC_012845	
		<i>Hydrometra</i> sp.	NC_012842	
		<i>Gerris</i> sp.	NC_012841	
		<i>Nerthra</i> sp.	NC_012838	
	Phthiraptera	<i>Bothriometopus macrocnemis</i>	NC_009983	
		<i>Campanulotes bidentatus</i>	NC_007884	
	Psocoptera	<i>Lepidopsocid</i> sp.	NC_004816	*
	Sternorrhyncha	<i>Acyrtosiphon pisum</i>	NC_011594	
		<i>Aleurochiton aceris</i>	NC_006160	
		<i>Aleurodicus dugesii</i>	NC_005939	
		<i>Bemisia tabaci</i>	NC_006279	
		<i>Neomaskellia andropogonis</i>	NC_006159	
		<i>Pachypsylla venusta</i>	NC_006157	
		<i>Schizaphis graminum</i>	NC_006158	
		<i>Tetraleurodes acaciae</i>	NC_006292	
		<i>Trialeurodes vaporariorum</i>	NC_006280	
		<i>Thrips imaginis</i>	NC_004371	
	Thysanoptera	<i>Lucanus mazama</i>	NC_013578	
		<i>Adelium</i> sp.	NC_013554	
	Coleoptera	<i>Chauliognathus opacus</i>	NC_013576	
		<i>Mordella atrata</i>	NC_013254	*
		<i>Rhopaea magnicornis</i>	NC_013252	
		<i>Macrogyrus oblongus</i>	NC_013249	
		<i>Psacotheta hilaris</i>	NC_013070	
		<i>Hydroscapha granulum</i>	NC_012144	
		<i>Trachypachus holmbergi</i>	NC_011329	
		<i>Anoplophora glabripennis</i>	NC_008221	
		<i>Tetraphalerus bruchi</i>	NC_011328	
		<i>Priasilpha obscura</i>	NC_011326	
		<i>Chaetosoma scaritides</i>	NC_011324	
		<i>Sphaerius</i> sp.	NC_011322	
		<i>Cyphon</i> sp.	NC_011320	
		<i>Crioceris duodecimpunctata</i>	NC_003372	
		<i>Pyrocoelia rufa</i>	NC_003970	
		<i>Pyrophorus divergens</i>	NC_009964	
		<i>Rhagophthalmus lufengensis</i>	NC_010969	
		<i>Rhagophthalmus ohbai</i>	NC_010964	
		<i>Tribolium castaneum</i>	NC_003081	
	Hymenoptera	<i>Acmaeodera</i> sp.	NC_013580	
		<i>Apis mellifera ligustica</i>	NC_001566	
		<i>Bombus ignitus</i>	NC_010967	
		<i>Vanhornia eucnemidarum</i>	NC_008323	
		<i>Evania appendigaster</i>	NC_013238	
		<i>Diadegma semiclausum</i>	NC_012708	
		<i>Orussus occidentalis</i>	NC_012689	
		<i>Cephus cinctus</i>	NC_012688	
		<i>Abispa ephippium</i>	NC_011520	
		<i>Adoxophyes honmai</i>	NC_008141	
	Lepidoptera	<i>Diatraea saccharalis</i>	NC_013274	
		<i>Lymantria dispar</i>	NC_012893	
		<i>Eriogyna pyretorum</i>	NC_012727	
		<i>Ochrogaster lunifer</i>	NC_011128	
		<i>Antheraea pernyi</i>	NC_004622	
		<i>Artogeia melete</i>	NC_010568	
		<i>Coreana raphaelis</i>	NC_007976	
		<i>Manduca sexta</i>	NC_010266	
		<i>Ostrinia furnacalis</i>	NC_003368	
		<i>Ostrinia nubilalis</i>	NC_003367	
		<i>Phthonandria atrilineata</i>	NC_010522	
		<i>Acraea issoria</i>	NC_013604	
		<i>Bombyx mandarina</i>	NC_003395	*

group	order	Species	GenBank accession no	reduced data set
	Mecoptera	<i>Neopanorpa pulchra</i>	NC_013180	
	Megaloptera	<i>Sialis hamata</i>	NC_013256	
		<i>Protohermes concolorus</i>	NC_011524	
		<i>Corydalus cornutus</i>	NC_011276	*
	Neuroptera	<i>Ditaxis biseriata</i>	NC_013257	
		<i>Ascaloptynx appendiculatus</i>	NC_011277	*
		<i>Polystoechotes punctatus</i>	NC_011278	
	Rhaphidioptera	<i>Mongoloraphidia harmandi</i>	NC_013251	
	Diptera (Brachycera)	<i>Chrysomya putoria</i>	NC_002697	
		<i>Cochliomyia hominivorax</i>	NC_002660	
		<i>Cydistomyia duplonotata</i>	NC_008756	
		<i>Dermatobia hominis</i>	NC_006378	
		<i>Drosophila mauritiana</i>	NC_005779	
		<i>Drosophila yakuba</i>	NC_001322	
		<i>Haematobia irritans irritans</i>	NC_007102	
		<i>Lucilia sericata</i>	NC_009733	
		<i>Simosyrphus grandicornis</i>	NC_008754	
		<i>Trichophthalma punctata</i>	NC_008755	
		<i>Hypoderma lineatum</i>	NC_013932	
		<i>Drosophila littoralis</i>	NC_011596	
		<i>Bactrocera carambolae</i>	NC_009772	
		<i>Bactrocera dorsalis</i>	NC_008748	
		<i>Bactrocera oleae</i>	NC_005333	
		<i>Bactrocera papayae</i>	NC_009770	
		<i>Bactrocera philippinensis</i>	NC_009771	
		<i>Ceratitis capitata</i>	NC_000857	
	Diptera (Nematocera)	<i>Aedes aegypti</i>	NC_010241	
		<i>Culicoides arakawae</i>	NC_009809	
		<i>Anopheles gambiae</i>	NC_002084	
		<i>Mayetiola destructor</i>	NC_013066	*
		<i>Rhopalomyia pomum</i>	NC_013063	
		<i>Anopheles quadrimaculatus</i>	NC_000875	

\* taxa included in reduced data set

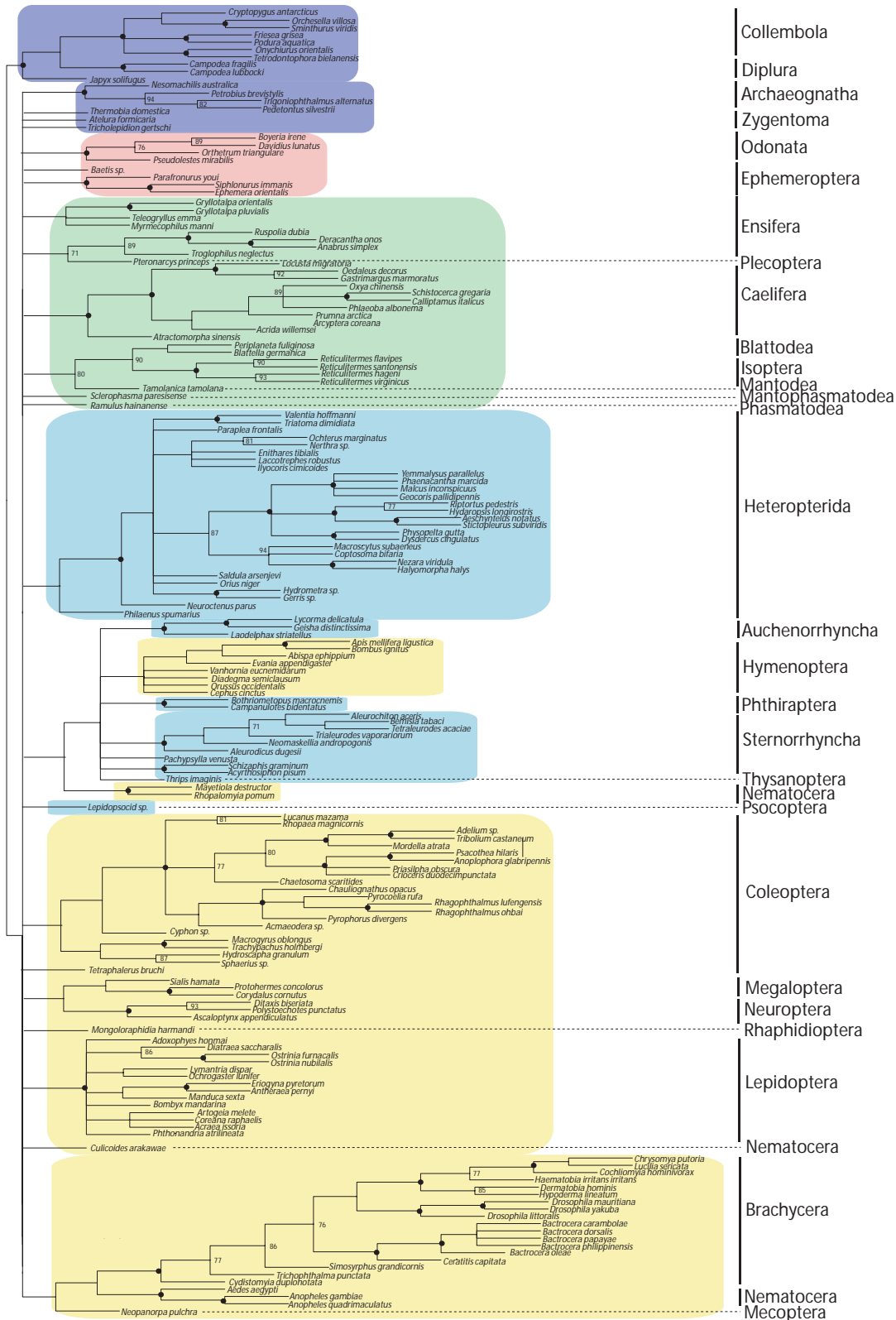
## Additional file 6

Randomly similar identified positions in the single mitochondrial gene alignments. Given are the percentages left after masking the single gene alignments (1<sup>st</sup>+2<sup>nd</sup> nucleotide positions of PCGs and amino acids) with ALICUT (<http://www.utilities.zfmk.de>) by excluding all randomly similar alignment positions.

	1 <sup>st</sup> +2 <sup>nd</sup> nt alignment	aa alignment
<i>atp6</i>		
Percent left:	89,87%	87,84%
<i>atp8</i>		
Percent left:	61,11%	53,98%
<i>cox1</i>		
Percent left:	89,75%	93,18%
<i>cox2</i>		
Percent left:	75,26%	89,90%
<i>cox3</i>		
Percent left:	97,23%	94%
<i>cytb</i>		
Percent left:	95,78%	97,04%
<i>nad1</i>		
Percent left:	92,49%	83,92%
<i>nad2</i>		
Percent left:	52,02%	56,36%
<i>nad3</i>		
Percent left:	82,10%	71%
<i>nad4</i>		
Percent left:	85,35%	70,57%
<i>nad4L</i>		
Percent left:	82,10%	48,54%
<i>nad5</i>		
Percent left:	80,25%	68,65%
<i>nad6</i>		
Percent left:	58,72%	36,33%

**Additional file 7**

MP topology (50% majority rule tree) inferred from the amino acid data set with all 174 mt genomes. The topology is based on 1000 bootstrap replicates with stepwise addition starting trees, simple addition of sequences and TBR branch-swapping using PAUP\*. Bootstrap values below 0.70 are not given, values above 0.95 are indicated by a dot. Primary wingless hexapods are indicated in blue, palaeopterous orders in red, polyneopterous orders in green, paraneopterous orders in turquoise and holometabolous orders in yellow.





### Additional file 8

Neighbor-Net graph based on split decomposition with the uncorrected  $p$  distance of the amino acid alignment of all 174 hexapods mtgenomes using SplitsTree4 after exclusion of randomly similar sections evaluated with ALISCORE. Primary wingless hexapods are indicated in blue, palaeopterous orders in red, polyneopterous orders in green, paraneopterous orders in turquoise and holometabolous orders in yellow.

